

POSTERIOR CONTRACTION IN SPARSE BAYESIAN FACTOR MODELS FOR MASSIVE COVARIANCE MATRICES

BY DEBDEEP PATI^{*,¶} ANIRBAN BHATTACHARYA^{†,||}, NATESH S.
PILLAI^{‡,**} AND DAVID DUNSON^{§,||}

*Department of Statistics, Florida State University[¶], Department of
Statistical Science, Duke University^{||}, Department of Statistics, Harvard
University^{**}*

Sparse Bayesian factor models are routinely implemented for parsimonious dependence modeling and dimensionality reduction in high-dimensional applications. We provide theoretical understanding of such Bayesian procedures in terms of posterior convergence rates in inferring high-dimensional covariance matrices where the dimension can be potentially larger than the sample size. Under relevant sparsity assumptions on the true covariance matrix, we show that commonly-used point mass mixture priors on the factor loadings lead to consistent estimation in the operator norm even when $p \gg n$. One of our major contributions is to develop a new class of continuous shrinkage priors and provide insights into their concentration around sparse vectors. Using such priors for the factor loadings, we obtain the same rate as obtained with point mass mixture priors. To obtain the convergence rates, we construct test functions to separate points in the space of high-dimensional covariance matrices using insights from random matrix theory; the tools developed may be of independent interest.

1. Introduction. It is now routine to collect data where the dimension p is much larger than the sample size n , and interest focuses on the covariance structure. In this context, even a simple parametric model like the Gaussian distribution leads to a high-dimensional model space, since an unstructured $p \times p$ covariance matrix has $O(p^2)$ free parameters. It is thus necessary to reduce the effective number of parameters via imposing sparsity or some lower-dimensional structure. Sparse Bayesian factor models ([West, 2003](#))

^{*}Assistant Professor, Department of Statistics, Florida State University

[†]Postdoctoral Associate, Department of Statistical Science, Duke University

[‡]Assistant Professor, Department of Statistics, Harvard University

[§]Professor, Department of Statistical Science, Duke University

AMS 2000 subject classifications: Primary 62G05, 62G20

Keywords and phrases: Bayesian estimation, Covariance matrix, Factor model, Rate of convergence, Shrinkage, Sparsity

provide one popular choice in applications, but currently lack theoretical support. In this paper, we close this gap by studying asymptotic properties for scenarios in which p grows faster than n .

Factor models (Bartholomew, 1987) aim to explain dependence among multivariate observations through shared dependence on a smaller number of latent factors. Given n i.i.d. observations $y_i \in \mathbb{R}^p$, the generic form of a latent factor model is

$$(1.1) \quad y_i = \mu + \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Omega), \quad i = 1, \dots, n,$$

where μ is an intercept term, Λ is a $p \times k$ factor loadings matrix with $k \ll p$, $\eta_i \sim N_k(0, I)$ are standard normal latent factors, and ϵ_i is a residual having diagonal covariance $\Omega = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$. We follow standard practice in centering the data prior to analysis and henceforth shall set $\mu = 0$ in (1.1). Marginalizing out the latent factors, $y_i \sim N_p(0, \Sigma)$ with

$$(1.2) \quad \Sigma = \Lambda \Lambda^T + \Omega,$$

which has at most $p(k+1)$ parameters, resulting in huge reduction in model complexity.

There is a sizeable literature studying asymptotic properties of various aspects of factor analysis, including consistent estimation of factor loadings and latent factors (Bai, 2003) and the number of factors (Bai and Ng, 2002; Lam and Yao, 2012). Fan, Fan and Lv (2008) studied rates of convergence of high-dimensional covariance estimates based on factor models, with Fan, Liao and Mincheva (2011) extending their results to approximate factor models that allow non-diagonal Ω in (1.2). This work assumes that the factor scores η_i are known, while we consider the fundamentally different setting in which the factor scores are unknown while also studying concentration of a Bayesian posterior instead of convergence of a point estimate.

A prior distribution on the loadings and the residual variances induces a prior distribution on the space of covariance matrices, and we are specifically interested in studying concentration of the corresponding posterior measure around the “true” covariance matrix. When the parameter space is finite dimensional, it is well known that the posterior contracts at the parametric rate of $n^{-1/2}$ under mild regularity conditions (Ghosal, Ghosh and van der Vaart, 2000). However, we are interested in the asymptotic framework of the dimension $p = p_n$ growing with the sample size n and hence the classical results do not apply to our case. Although this setting has motivated abundant frequentist work, relatively little has been done in the Bayesian setting, with most of the focus being on linear regression and the closely-related normal means problem; relevant references include Armagan, Dunson and Lee

(2011); Belitser and Ghosal (2003); Bontemps (2011); Castillo and van der Vaart (2012); Ghosal (1997) among others. In fact, to best of our knowledge, ours is the first paper which studies the asymptotic properties of Bayesian covariance estimation in this context.

Now we summarize the main results obtained in this paper. We begin with the study of a *moderately-high-dimensional* setting where p_n grows slower than n . A Bayesian specification (Arminger and Muthén, 1998; Song and Lee, 2001) of the factor model for such applications commonly uses inverse gamma priors on the residual variances and normal prior on the loadings. We show in Theorem 3.3 below that such priors lead to a posterior contraction rate of $\sqrt{p_n^\gamma/n}$ in the Frobenius norm whenever the true covariance underlying the data generating mechanism admits a factor decomposition as in (1.2) with the number of factors $k = k_n = O(1)$. Thus even if p_n is allowed to grow with n , we obtain posterior consistency as long as $p_n^\gamma/n \rightarrow 0$ for some $\gamma \geq 1$.

The second set of results pertain to the more interesting case, $p_n \gg n$. In this regime, we shall consider a weaker notion of discrepancy, namely the operator norm or equivalently the largest eigenvalue. Although the original specification of the factor model reduces the number of variables from $O(p_n^2)$ to $O(p_n)$, the estimation problem is still challenging when $p_n \gg n$. To address this challenge, West (2003) introduced *sparse factor modeling* to allow many of the loadings to be exactly equal to zero through a point mass mixture prior; see also Carvalho et al. (2008); Lucas et al. (2006). We show in Theorem 3.5 that, for appropriate point mass priors, the posterior distribution contracts at a rate of $O(\sqrt{(\log p_n)^\gamma/n})$ in the operator norm if the true sequence of covariance matrices admit a factor type decomposition (1.2) with $k_n = O(1)$ many factors and some realistic sparsity assumption on the loadings. Thus, we obtain consistency as long as $p_n = O(e^{n^\alpha})$ for some $\alpha \in (0, 1/\gamma)$. This is particularly appealing since the dimensionality affects the rate only through a logarithmic factor and thus provides a theoretical validation of the efficacy of sparse factor models for high-dimensional covariance matrix estimation for $p_n \gg n$.

Our final set of results concern with developing continuous shrinkage priors that achieve the same rate of convergence as that of the point mass mixture priors. Although point mass mixture priors on factor loadings are intuitively appealing and possess attractive theoretical properties, they lead to daunting posterior computation, with typical MCMC algorithms for updating elements of the loadings matrix one at a time facing problems with slow convergence and mixing. To address such problems through block updating, while allowing a weaker notion of sparsity in which elements are close to zero instead of exactly zero, continuous shrinkage priors can be used. Such

priors have become common in regression (Armagan, Dunson and Lee, 2011; Carvalho, Polson and Scott, 2010; Hans, 2011; Park and Casella, 2008), with Polson and Scott (2010) providing a unifying local-global scale mixture representation. The lack of tight concentration bounds for such priors has limited the study of their asymptotic properties. We develop a novel class of local-global shrinkage priors for which such bounds can be obtained, leading to a rate of $\sqrt{(\log p_n)^\gamma/n}$ in operator norm in the $p_n \gg n$ setting.

Technically, our methods proceed via the usual route of first establishing the prior concentration and constructing test functions with appropriate rates independently and then combining these two using entropy calculations. For constructing test functions for the operator norm, the traditional tests based on likelihood ratios do not yield the right rates. Instead we construct tests inspired by results from the non-asymptotic theory of random matrices and hence these calculations are of independent interest and may be useful to other problems in high-dimensional estimation.

High-dimensional covariance matrix estimation has been widely studied from a frequentist perspective. The inadequacy of the sample covariance in $p_n \gg n$ settings is well known, motivating regularized estimators based on banding or tapering the sample covariance matrix (Bickel and Levina, 2008b; Furrer and Bengtsson, 2007; Wu and Pourahmadi, 2010), banding the Cholesky factor (Wu and Pourahmadi, 2003), regularizing the inverse Cholesky factor (Huang et al., 2006; Levina, Rothman and Zhu, 2008), thresholding the sample covariance matrix (Bickel and Levina, 2008a; Cai and Liu, 2011; El Karoui, 2008), regularizing the precision matrix (Rothman et al., 2008) and regularized principal component analysis (Johnstone and Lu, 2009; Zou, Hastie and Tibshirani, 2006) among others. Theoretical properties of such regularized estimators have been studied in Bickel and Levina (2008a,b); El Karoui (2008); Lam and Fan (2009), with explicit rates of convergence obtained in an asymptotic framework where p_n increases with n . Minimax optimal rates in operator & Frobenius norm have also been recently established in Cai, Zhang and Zhou (2010).

The rest of the paper is organized as follows. After setting up the basic notations and definitions in Section 2, we present the main results of this paper in Section 3. In Section 4, we discuss and provide guidelines for prior elicitation based on the theoretical results. Section 5 develops a number of auxiliary results of independent interest that are used to prove the main results in Section 6. The proof of some technical lemmas are given in an Appendix.

2. Preliminaries. Given sequences a_n, b_n , we shall denote $a_n = O(b_n)$ if there exists a global constant C such that $a_n \leq Cb_n$.

Given a metric space (X, d) , let $N(\epsilon; X, d)$ denote its ϵ -covering number, i.e., the minimum number of balls of radius ϵ_n needed to cover X .

For a vector $x \in \mathbb{R}^r$, $\|x\|_2$ denotes its Euclidean norm. We will use \mathcal{S}^{r-1} to denote the unit Euclidean sphere $\{x \in \mathbb{R}^r : \|x\|_2 = 1\}$ and Δ^{r-1} to denote the $(r-1)$ -dimensional simplex $\{x = (x_1, \dots, x_r)^\top : x_j \geq 0, \sum_{j=1}^r x_j = 1\}$. Further, let Δ_0^{r-1} denote $\{x = (x_1, \dots, x_{r-1})^\top : x_j \geq 0, \sum_{j=1}^{r-1} x_j \leq 1\}$.

For a square matrix A , $\text{tr}(A)$ and $|A|$ respectively denote the trace and the determinant of A . For a $p \times r$ matrix $A = (a_{jj'})$ with $p \geq r$, using the singular value decomposition we may write

$$A = \sum_{k=1}^r s_{(k)} u_k v_k'$$

where $s_{(1)} \geq s_{(2)} \geq \dots \geq s_{(r)} \geq 0$ denote the singular values of A (or equivalently the eigenvalues of $\sqrt{A^\top A}$) arranged in decreasing order and u_k, v_k denote the corresponding singular vectors. We shall also use $s_{\min}(A)$ and $s_{\max}(A)$ to denote the smallest and largest singular values respectively. We will investigate the posterior convergence rates for two norms; the Frobenius norm ($\|\cdot\|_F$) and the operator norm ($\|\cdot\|_2$) defined in the usual way:

$$\begin{aligned} \|A\|_F &= \sqrt{\sum_{j=1}^p \sum_{j'=1}^r a_{jj'}^2} = \sqrt{\text{tr}(A^\top A)} \\ \|A\|_2 &= \sup_{x \in \mathcal{S}^{r-1}} \|Ax\|_2 = s_{\max}(A). \end{aligned}$$

Clearly, for any fixed dimension p , the above two norms are equivalent and thus convergence rate in one norm will lead to identical convergence in the other. However this is no longer the case when the dimension $p = p_n$ grows with n . In fact we will see below that the convergence rates are indeed different for the two norms above.

For a subset $S \subset \{1, \dots, p\}$, let $|S|$ denote the cardinality of S and define $\theta_S = (\theta_j : j \in S)$ for a vector $\theta \in \mathbb{R}^p$. Denote $\text{supp}(\theta)$ to be the *support* of θ , i.e., the subset $S_0 \subset \{1, \dots, p\}$ corresponding to the non-zero entries of θ . We shall continue to use the same notations for a subset of entries and support for matrices Λ , where it has to be interpreted that Λ is vectorized column-wise. Let $l_0[s; p]$ denote the subset of \mathbb{R}^p given by

$$l_0[s; p] = \{x \in \mathbb{R}^p : \#(1 \leq j \leq p : x_j \neq 0) \leq s\}.$$

Clearly, $l_0[s; p]$ consists of s -sparse vectors θ with $|\text{supp}(\theta)| \leq s$.

$\text{IG}(a, b)$ (resp. $\text{IG}_{[c, d]}(a, b)$) denotes the inverse-gamma density with parameters a, b (resp. truncated to the interval $[c, d]$). Throughout C, C' are generically used to denote positive constants which are irrelevant to our purpose.

Finally, let \mathcal{C}_n denote the cone of covariance matrices of size $p_n \times p_n$ and let $\Sigma_{0n} \in \mathcal{C}_n$ denote a true sequence of covariance matrices. We observe

$$y_1, \dots, y_n \stackrel{\text{i.i.d.}}{\sim} N_{p_n}(0, \Sigma_{0n})$$

and set $\mathbf{y}^{(n)} = (y_1, \dots, y_n)$. We model the data as

$$(2.1) \quad y_i \stackrel{\text{i.i.d.}}{\sim} N_{p_n}(0, \Sigma_n), \quad \Sigma_n = \Lambda_n \Lambda_n^\top + \Omega_n.$$

3. Main results. In this section we present the main results of this paper. Let $\Theta_\Lambda^{(p, k)} \in \mathbb{R}^{p \times k}$ denote the class of real-valued $p \times k$ matrices. We start with the following assumptions on the true covariance matrix of the observed data $\mathbf{y}^{(n)}$.

ASSUMPTION 3.1. *The true sequence of covariance matrices Σ_{0n} are of the form*

$$(A0) \quad \Sigma_{0n} = \Lambda_{0n} \Lambda_{0n}^\top + \Omega_{0n}, \quad \Lambda_{0n} \in \Theta_\Lambda^{(p_n, k_{0n})}, \quad \Omega_{0n} = \sigma_{0n}^2 \mathbf{I}_{p_n},$$

and $k_{0n} = O(1)$ is known.

Assumption (3.1) says that the true sequence of covariances Σ_{0n} admit a factor decomposition as in (1.2) with $\Omega_{0n} = \sigma_{0n}^2 \mathbf{I}_{p_n}$. We assume $\Omega_n = \sigma^2 \mathbf{I}_{p_n}$ in (2.1). We also assume the true sequence of factors k_{0n} to be bounded and known for notational simplicity and it is to be understood that the model (2.1) is fitted with $k_n = k_{0n}$ many factors. However, we have identified the role of k_n in all calculations and our results can be easily relaxed by placing a prior on the number of factors k_n and assuming an appropriate growth of the true number of factors k_{0n} .

A prior distribution $\Pi_n(\Lambda_n \otimes \sigma^2)$ on $\Theta_\Lambda^{(p_n, k_n)} \times \mathbb{R}^+$ induces a prior distribution on \mathcal{C}_n , which is also denoted by $\Pi_n(\Sigma_n)$. We will denote by $\Pi_n(\cdot | \mathbf{y}^{(n)})$ the corresponding posterior distribution for Σ_n . For a sequence of numbers $\epsilon_n \rightarrow 0$ and a constant $M > 0$ independent of ϵ_n (to be chosen later), let

$$(3.1) \quad U_n = \{\Sigma_n : \|\Sigma_n - \Sigma_{0n}\| \leq M\epsilon_n\}$$

denote a ball of radius $M\epsilon_n$ around Σ_{0n} with respect to some matrix norm $\|\cdot\|$; we shall focus on the Frobenius norm ($\|\cdot\|_F$) and the operator norm

($\|\cdot\|_2$) in the sequel. For various prior distributions on $\Theta_{\Lambda}^{(p_n, k_n)} \times \mathbb{R}^+$, we seek to find a minimum such possible sequence ϵ_n and a subset $\mathcal{C}_{0n} \subset \mathcal{C}_n$ such that for any $\Sigma_{0n} \in \mathcal{C}_{0n}$,

$$(3.2) \quad \lim_{n \rightarrow \infty} \Pi_n(U_n^c \mid \mathbf{y}^{(n)}) = 0, \quad \text{in probability}$$

where $\Pi_n(U_n^c \mid \mathbf{y}^{(n)})$ denotes the posterior probability of the event U_n^c . Notice that the posterior measure is random since it is conditioned on the observed data; thus the above limit is over the probability space corresponding to the observed data.

3.1. Frobenius norm. We now mention specific assumptions on Σ_{0n} and prior choices.

ASSUMPTION 3.2. *In addition to Assumption 3.1, the true covariance matrix $\Sigma_{0n} \in \mathcal{C}_n$ satisfies the following:*

(AF1) $p_n < n$ with $\lim_{n \rightarrow \infty} p_n^\gamma / n \rightarrow 0$ for some $\gamma \geq 1$.

(AF2) $\frac{1}{\log n} \leq \sigma_{0n}^2 \leq M_\sigma$.

We assume the following prior on $[\Lambda_n]_{jh} = \lambda_{jh}$ and σ^2 ,

$$(P0) \quad \lambda_{jh} \sim N(0, 1), j = 1, \dots, p_n, h = 1, \dots, k_n, \sigma^2 \sim \text{IG}_{[0, M_\sigma]}(a, b).$$

We now state our main theorem on posterior convergence rates in Frobenius norm.

THEOREM 3.3. *Suppose the true sequence of covariance matrices Σ_{0n} satisfy Assumption 3.2 with $\gamma \geq 9$ in **(AF1)**. Also, assume the prior distribution $\Pi_n(\Lambda_n \otimes \sigma^2)$ as in **(P0)**. Then with $\epsilon_n = \sqrt{\frac{p_n^9}{n} (\log n)^3}$ and for some $M > 0$ large enough,*

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\Sigma_{0n}} \Pi_n(\|\Sigma_n - \Sigma_{0n}\|_F > M\epsilon_n \mid \mathbf{y}^{(n)}) = 0.$$

3.2. Operator norm. [Cai, Zhang and Zhou \(2010\)](#) showed that the min-max optimal rate in operator norm is given by $\sqrt{\log p_n / n}$ for their sparsity class. Although their sparsity class is different from the one implied by factor models, it would be appealing to obtain a similar rate of convergence as the effect of dimensionality enters only through a logarithmic factor. We now mention specific assumptions on Σ_{0n} and prior choices.

ASSUMPTION 3.4. *In addition to **(A0)** in Assumption 3.1, the true covariance matrix $\Sigma_{0n} \in \mathcal{C}_n$ satisfies the following:*

- (A1) $\lim_{n \rightarrow \infty} (\log p_n)^\gamma / n = 0$ for some $\gamma \geq 1$.
 (A2) Each column of Λ_{0n} belongs to $l_0[s_n; p_n]$ with $s_n = O(\log p_n)$.
 (A3) There exists a sequence of positive real numbers c_n with $c_n = O(\log p_n)$ such that

$$\left\| \frac{1}{c_n} \Lambda_{0n}^\top \Lambda_{0n} - \mathbf{I}_{k_n} \right\|_2 = O(\sqrt{k_n/p_n}).$$

- (A4) There exist constants $\sigma_0^{(1)}$ and $\sigma_0^{(2)}$ such that $\sigma_0^{(1)} \leq \sigma \leq \sigma_0^{(2)}$.

We now discuss implications of each of the above assumptions.

- Assumption (A1) allows p_n to grow faster than n under the very mild assumption of $(\log p_n)^\gamma / n \rightarrow 0$. In particular, p_n can be of the order of $\exp(n^\alpha)$ for any $\alpha \in (0, 1/\gamma)$.
- Following the motivation in West (2003), one requires sparsity in the loadings for meaningful inference in $p_n \gg n$ situations. This is reflected through (A2), requiring the loadings columns to be sparse with $O(\log p_n)$ many signals per column. Notice that this is where we differ from sparsity assumptions used by previous authors (Bickel and Levina, 2008a; Cai and Liu, 2011; Cai, Zhang and Zhou, 2010; Levina, Rothman and Zhu, 2008). Even if the entries of Λ are exactly zero, the corresponding covariance matrix need not have many zero entries.
- Conditions similar to (A3) have been used previously in the econometric factor model setting (Fan, Fan and Lv, 2008; Fan, Liao and Mincheva, 2011) and referred to as “pervasive”, meaning the factors influence all the variables. We provide a different intuition based on random matrix theory which suggests that (A3) is indeed mild and expected to be satisfied by a large class of loadings.

If the elements of the $p_n \times k_n$ matrix Λ_{0n} are drawn i.i.d. from a $N(0, 1)$ distribution, then Theorem 5.39 of Vershynin (2010) tells us that

$$\left\| \frac{1}{p_n} \Lambda_{0n}^\top \Lambda_{0n} - \mathbf{I}_{k_n} \right\|_2 \leq C \frac{\sqrt{k_n}}{\sqrt{p_n}}$$

with probability at least $1 - e^{-C' k_n}$ for some constants $C', C > 0$. Equivalently, all singular values of $\Lambda_{0n}/\sqrt{p_n}$ lie in $(1 - C \frac{\sqrt{k_n}}{\sqrt{p_n}}, 1 + C \frac{\sqrt{k_n}}{\sqrt{p_n}})$ with high probability. Intuitively, this tells us that “tall and skinny” matrices when appropriately normalized behave as approximate isometries.

As our emphasis is on sparse factor models, a more realistic generative model for the loadings would be

$$\lambda_{0jh} \sim (1 - \pi_n) \delta_0 + \pi_n N(0, 1),$$

where $\lambda_{0jh} = [\Lambda_{0n}]_{jh}$, δ_0 denotes a point mass at zero and $\pi_n = s_n/p_n$ to reflect the sparsity assumption in **(A2)**.

A modification of Theorem 5.39 of [Vershynin \(2010\)](#) implies

$$\left\| \frac{1}{p_n} \Lambda_{0n}^\top \Lambda_{0n} - \pi_n \mathbf{I}_{k_n} \right\|_2 \leq C \frac{\sqrt{k_n}}{\sqrt{p_n}} \|\pi_n \mathbf{I}_{k_n}\|_2,$$

which in turn yields that

$$\left\| \frac{1}{s_n} \Lambda_{0n}^\top \Lambda_{0n} - \mathbf{I}_{k_n} \right\|_2 \leq C \frac{\sqrt{k_n}}{\sqrt{p_n}},$$

with probability at least $1 - e^{-C'k_n}$. Since $s_n = O(\log p_n)$ by **(A2)**, we let the normalizer c_n in **(A3)** to be $O(\log p_n)$.

- **(A4)** simply posits that the residual variance is bounded above and below. The lower bound is used to avoid Σ_{0n} from being ill-conditioned. See Remark 3.7 for a discussion on relaxing this assumption.

We now define our prior $\Pi_n(\Lambda \otimes \sigma^2)$ on $\Theta_\Lambda^{(p_n, k_n)} \times \mathbb{R}^+$ through independent priors on the loadings Λ and the residual variance σ^2 . We draw σ^2 from a density f_σ on $(0, \infty)$,

$$(PR) \quad \sigma^2 \sim f_\sigma(\cdot).$$

We first consider a class of point mass mixture priors on the loadings similar to that advocated by [West \(2003\)](#),

$$(PL1) \quad \begin{aligned} \lambda_{jh} &\sim (1 - \pi)\delta_0 + \pi g(\cdot), \quad j = 1, \dots, p_n; \quad h = 1, \dots, k_n, \\ \pi &\sim \text{Beta}(1, \kappa p_n + 1), \quad \kappa > 0, \end{aligned}$$

where δ_0 denotes a point mass at zero and g is an absolutely continuous density on \mathbb{R} with exponential tails or heavier.

In the context of linear regression, [Scott and Berger \(2010\)](#) showed that such point mass mixture priors with a beta hyper-prior on the mixture probability lead to an automatic multiplicity correction. [Jiang \(2007\)](#) proved optimality results in estimating the predictive under such priors in generalized linear models accommodating diverging numbers of predictors. [Castillo and van der Vaart \(2012\)](#) studied concentration properties of a class of prior distributions similar to **(PL1)** on a high-dimensional normal mean and showed that they lead to the minimax optimal rate of convergence.

With the prior specification complete, we are in a position to state the first theorem on posterior contraction rates in the operator norm.

THEOREM 3.5. *Suppose the true covariance matrix $\Sigma_{0n} \in \mathcal{C}_n$ satisfies Assumptions **(A0)** – **(A4)** in Assumption 3.4 with $\gamma \geq 5$ in **(A1)**. Also assume independent priors $\Pi(\Lambda)$ and $\Pi(\sigma^2)$ on the loadings and the residual variances as in (PL1) and (PR) respectively with f_σ being a gamma(a, b) density. Then with $\epsilon_n = \sqrt{\frac{(\log p_n)^5}{n}}$,*

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\Sigma_{0n}} \Pi_n(\|\Sigma_n - \Sigma_{0n}\|_2 > M\epsilon_n \mid \mathbf{y}^{(n)}) = 0.$$

As mentioned in the Introduction, although point mass mixture priors are conceptually appealing in allowing exact sparsity and often leading to appealing theoretical properties, posterior computation under such priors is extremely daunting computationally in high-dimensional cases. As an alternative, a rich variety of continuous shrinkage priors have been developed that admit a scale mixture representation (Polson and Scott, 2010). A fundamental hurdle in studying theoretical properties of such priors is the difficulty of obtaining tight bounds on their concentration. With the motivation of developing a continuous shrinkage prior that can be shown to concentrate on sparse vectors and approximate point mass mixture priors, we propose a novel class of priors. We use such priors for the factor loadings, but they should be broadly applicable in other high-dimensional settings.

Let $\text{DE}(\psi)$ denote the Laplace or double-exponential density with scale parameter ψ with a density given by

$$(3.3) \quad f(x) = \frac{1}{2\psi} e^{-\frac{|x|}{\psi}}, \quad x \in \mathbb{R}.$$

Draw the elements of a high-dimensional vector $\theta \in \mathbb{R}^p$ through the following hierarchical mechanism:

$$(PS) \quad \theta_j \sim \text{DE}(\tau\gamma_j), \quad \tau \sim f_\tau, \quad \gamma \sim f_\gamma,$$

where f_τ and f_γ are densities on \mathbb{R}^+ and Δ_0^{p-1} respectively. In particular, we require f_τ to satisfy (a) $\mathbb{P}(\tau > \log p) \leq e^{-C \log p}$, (b) $\mathbb{P}(\tau \in [2 \log p, 4 \log p]) \geq e^{-C \log p}$ and (c) $\mathbb{P}(\tau < 1/\log p) \leq e^{-C \log p}$ for large values of p . In Lemma A.1 stated in the Appendix, we show that the $\text{IG}(\log p, \log p)$ distribution is one possible candidate for f_τ . We also choose f_γ to be a $\text{Dir}(\alpha/p, \dots, \alpha/p)$ density, where $\text{Dir}(\alpha_1, \dots, \alpha_p)$ denotes a Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_p$ which has a density f_γ on Δ_0^{p-1} given by

$$f_\gamma(x_1, \dots, x_{p-1}) = \frac{\Gamma(\alpha)}{\prod_{j=1}^p \Gamma(\alpha_j)} \prod_{j=1}^{p-1} x_j^{\alpha_j-1} \left(1 - \sum_{j=1}^{p-1} x_j\right)^{\alpha_p-1}.$$

Although the prior specification in (PS) has similarities to local-global shrinkage rules in Polson and Scott (2010), a main difference is that the local scale parameters in γ are drawn jointly from a Dirichlet distribution instead of independent draws from a continuous distribution on \mathbb{R}^+ . In particular, constraining the local scale parameters to lie on the simplex prevents the l_1 norm of θ from blowing up with increasing dimension, while a $\text{Dir}(\alpha/p, \dots, \alpha/p)$ prior on γ ensures that a handful of entries are left unshrunk with the rest heavily shrunk towards zero. For a detailed discussion on our proposed prior and connections to point mass mixture priors, refer to Section 4.

We show in Theorem 3.6 that our proposed shrinkage prior on the vectorized loadings indeed works as a surrogate to the point mass mixture priors because they achieve the same posterior rate of convergence (up to a log factor) as in Theorem 3.5.

THEOREM 3.6. *Suppose the true covariance matrix $\Sigma_{0n} \in \mathcal{C}_n$ satisfies (A0) – (A4) in Assumption 3.4 with $\gamma \geq 5$ in (A1). Furthermore, suppose that the vectorized loadings are drawn according to the shrinkage prior in (PS) and the prior on σ^2 is as in (PR) with f_σ being a gamma(a, b) density. Then, with $\epsilon_n = \sqrt{\frac{(\log p_n)^5}{n}}$,*

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\Sigma_{0n}} \Pi_n(\|\Sigma_n - \Sigma_{0n}\|_2 > M\epsilon_n \mid \mathbf{y}^{(n)}) = 0.$$

The following remark clarifies the lower bound assumption on σ in (A4).

REMARK 3.7. (A4) assumes the lower bound $\sigma_0^{(1)}$ to be fixed rather than decaying with p_n for technical simplicity. We claim without proof that one can actually let $\sigma_0^{(1)} = C/(\log p_n)^{1/4}$ incurring only minor changes in the proofs for Theorems 3.5 and 3.6. In that case the rates of convergence are slowed down by a $(\log n)$ term, i.e., $\epsilon_n = \sqrt{\frac{(\log p_n)^5}{n}}(\log n)^\kappa$ for some constant $\kappa > 0$.

4. Shrinkage prior in high-dimensional settings. Let θ be a p -dimensional vector and $\theta_0 \in l_0[s; p]$ be an s -sparse vector with $s = O(\log p)$. Depending on the problem, θ might correspond to a high-dimensional mean vector, a vector of regression coefficients or a column of the factor loadings, with θ_0 corresponding to a sparse truth. A quantity of fundamental importance in studying the behavior of the posterior distribution in these high-dimensional problems is the prior concentration or the non-centered

small ball probability

$$(4.1) \quad \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon)$$

around sparse vectors θ_0 . It can be shown that if θ_j 's are i.i.d. standard normal,

$$\sup_{\theta_0 \in l_0[s;p]} \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon) \leq e^{-Cp \log(\frac{1}{\epsilon})}$$

which decays exponentially with p for fixed s limiting the ability of the posterior to concentrate on sparse θ_0 .

However, with appropriate point mass mixture priors, the small ball probability (4.1) can be improved to $e^{-Cs \log(\frac{1}{\epsilon})}$. We thus discuss some of the salient features of point mass mixture priors here and illustrate how these features can give insights for developing continuous shrinkage priors.

Castillo and van der Vaart (2012) recommended the following hierarchical prior on θ :

- (P1) An integer j is chosen according to a prior probability π_p on $\{1, \dots, p\}$.
- (P2) A subset S of size j is chosen uniformly at random from the $\binom{p}{j}$ subsets of size j .
- (P3) Given (j, S) , elements of θ_S are drawn independently from a probability distribution with Lebesgue measure g on \mathbb{R} and this is extended to $\theta \in \mathbb{R}^p$ by setting the remaining coordinates to 0.

The commonly-used point mass mixture priors of the form $\theta_j \sim (1 - \pi)\delta_0 + \pi g$ arise a special case of the above general framework with the prior π_p on the subset size corresponding to the Binomial(p, π) prior.

Now suppose $\theta_0 \in l_0[s;p]$ and let S_0 denote the support of θ_0 . Then,

$$(4.2) \quad \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon) \geq \Pi(S_0) \mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\|_2 < \epsilon),$$

where $\Pi(S_0)$ denotes the prior probability of choosing the subset S_0 . In particular, under the Binomial(p, π) prior on the subset size, we have

$$\Pi(S_0) = \pi^s (1 - \pi)^{p-s}.$$

If one knew s beforehand, an intuitive choice for π is s/p , with the corresponding prior referred to as the *oracle prior* by Castillo and van der Vaart (2012). With this choice,

$$\Pi(S_0) = \left(\frac{s}{p}\right)^s \left(1 - \frac{s}{p}\right)^{p-s} = \left(\frac{s}{p}\right)^s \left(1 - \frac{s}{p}\right)^{\frac{p}{s}s(1-s/p)} \geq e^{s \log s - s \log p + s/4}$$

and

$$\mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\| < \epsilon) \geq e^{-Cs \log(1/\epsilon)}$$

leading to a higher prior concentration. [Castillo and van der Vaart \(2012\)](#) further showed that even without knowledge of s , one can achieve similar concentration around sparse vectors through a Beta hyperprior $\pi \sim \text{Beta}(1, \kappa p + 1)$, $\kappa > 0$.

To avoid the computational difficulties associated with point mass mixture priors, a number of recent works aim to develop a continuous shrinkage prior that effectively mimics the mixture priors. [Polson and Scott \(2010\)](#) unified a number of such priors through the following scale-mixture representation:

$$(4.3) \quad \theta_j \sim N(0, \psi_j \phi), \quad \psi_j \sim \pi(\psi_j), \quad \phi \sim \pi(\phi),$$

where ψ_j and ϕ are local and global scale parameters, respectively. Despite computational advantages with this family of shrinkage priors, it is not clear whether they have adequate concentration around s -sparse vectors. We found that a suitable dependence structure in (ψ_1, \dots, ψ_p) can force a large subset of the local scales ψ_j to be simultaneously close to zero and thus achieve a concentration similar to point mass mixture priors. This observation motivated the shrinkage prior (PS) in Section 3, where we let $\psi_j = \tau \gamma_j$ with $\tau > 0$ and $\gamma = (\gamma_1, \dots, \gamma_p)^T \in \Delta^{p-1}$ with $\gamma \sim \text{Dir}(\alpha/p, \dots, \alpha/p)$.

We now exhibit some aspects of our proposed shrinkage prior (PS). We shall first show that (PS) achieves the same concentration around sparse vectors as the point mass mixture priors in [Castillo and van der Vaart \(2012\)](#). We further exhibit a tail bound on the number of “large signals” implied by (PS) and conclude the section by proving a large deviation result for the l_1 norm of a vector drawn from (PS).

In the following Lemma 4.1, we show that under a mild restriction on the magnitude of the non-zero entries of θ_0 , the hierarchical prior specification in (PS) leads to the same order of concentration around elements in $l_0[s; p]$ as (P1) – (P3).

LEMMA 4.1. *Suppose θ is drawn according to the prior (PS). Let $\theta_0 \in l_0[s; p]$, $1 \leq s \leq p$ with $\|\theta_0\|_1 = O(s \log s)$ and $s/p \leq 1/2$. Then, for any $\epsilon \in (0, 1)$,*

$$\mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon) \geq \exp[-C \max\{s \log(s/\epsilon), \log p\}]$$

for some constant $C > 0$.

PROOF. Let $\delta = \epsilon/p$. To lower-bound $\mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon)$, we first obtain a lower bound conditioned on the hyper parameters τ and γ :

$$\begin{aligned}
 & \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon \mid \tau, \gamma) \\
 & \geq \mathbb{P}(|\theta_j| \leq \delta \mid \tau, \gamma) \mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\|_2 < \epsilon/2 \mid \tau, \gamma) \\
 (4.4) \quad & = \left[\prod_{j \in S_0^c} \left(1 - e^{-\frac{\delta}{\psi_j}} \right) \right] \times \mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\|_2 < \epsilon/2 \mid \tau, \gamma).
 \end{aligned}$$

Let $\tilde{\gamma} = (\gamma_1, \dots, \gamma_{p-1})^T$ and $\gamma_p = 1 - \sum_{j=1}^{p-1} \gamma_j$. We now have to integrate out τ and $\tilde{\gamma}$ in (4.4). By a relabeling of indices, we can always make sure that the p th index lies in S_0 . Let $S_1 = S_0 \setminus \{p\}$ so that $S_0^c \cup S_1 = \{1, \dots, p-1\}$. For a fixed τ in the interval $[2s, 4s]$ and numbers $a, b \in (0, 1)$ with $b = 4a$, let \mathcal{A}_τ denote the subset of Δ_0^{p-1} given by,

$$(4.5) \quad \mathcal{A}_\tau = \left\{ 0 \leq \gamma_j \leq \frac{\delta}{\log(p/s)\tau} \mid j \in S_0^c; \gamma_j \in \left[\frac{a}{\tau}, \frac{b}{\tau} \right] \mid j \in S_1 \right\}.$$

Observe that \mathcal{A}_τ defines a valid subset of Δ_0^{p-1} for ϵ small enough, since $\gamma_j \geq 0$ for all $j = 1, \dots, p-1$ and

$$(4.6) \quad \sum_{j=1}^{p-1} \gamma_j = \sum_{j \in S_0^c} \gamma_j + \sum_{j \in S_1} \gamma_j \leq \epsilon + \frac{(s-1)b}{2s} \leq b < 1$$

for $\epsilon < b/2$. Thus,

$$\begin{aligned}
 \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon) &= \int_{(\tau, \tilde{\gamma}) \in \mathbb{R}^+ \times \Delta_0^{p-1}} \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon \mid \tau, \tilde{\gamma}) f_\gamma(d\tilde{\gamma}) f_\tau(d\tau) \\
 (4.7) \quad &\geq \int_{(\tau, \gamma) \in \mathcal{B}} \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon \mid \tau, \gamma) f_\gamma(d\tilde{\gamma}) f_\tau(d\tau)
 \end{aligned}$$

where $\mathcal{B} = \cup_{\tau \in [2s, 4s]} \mathcal{B}_\tau$ with $\mathcal{B}_\tau = \{\tau\} \times \mathcal{A}_\tau \subset \mathbb{R}^+ \times \Delta_0^{p-1}$. We now substitute the lower bound for $\mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon \mid \tau, \gamma)$ from (4.4) in (4.7) and lower-bound the two terms on the right hand side of (4.4) individually.

For the first term, observe that for $(\tau, \gamma) \in \mathcal{B}$,

$$\prod_{j \in S_0^c} \left(1 - e^{-\frac{\delta}{\psi_j}} \right) \geq (1 - s/p)^{p-s}.$$

To tackle the second term, we make use of the following Lemma 4.2 whose proof is provided in the Appendix.

LEMMA 4.2. *Let $\eta \in \mathbb{R}^s$ denote a random vector with independent components $\eta_j \sim DE(\psi_j)$. If there exist numbers $a, b > 0$ such that $\psi_j \in [a, b]$ for all $j = 1, \dots, s$, then for any $\delta > 0$ and $\eta_0 \in \mathbb{R}^s$,*

$$\mathbb{P}(\|\eta - \eta_0\|_2 < \delta) \geq \exp \left\{ -s \log 2 - \sum_{j=1}^s \frac{|\eta_{0j}|}{a} \right\} \left(1 - e^{-\delta/(b\sqrt{s})} \right)^s.$$

By definition, $\psi_j \in [a, b]$ for all $j \in S_1$ whenever $(\tau, \gamma) \in \mathcal{B}$. Further, along the lines of (4.6), $\sum_{j=1}^{p-1} \gamma_j \in [a/8, b]$ and hence $\gamma_p \in [1-b, 1-a/8]$ on \mathcal{B} . Since a, b are constants, by a slight abuse of notation, we shall assume $\psi_j \in [a, b]$ for all $j \in S_0$ on \mathcal{B} . It thus follows from Lemma 4.2 that

$$\begin{aligned} & \mathbb{P}(\|\Pi_{S_0}(\theta) - \Pi_{S_0}(\theta_0)\|_2 < \epsilon/2 \mid \tau, \gamma) \\ & \geq \exp \left\{ -s \log 2 - \sum_{j \in S_0} \frac{|\theta_{0j}|}{a} \right\} \left(1 - e^{-\epsilon/(2b\sqrt{s})} \right)^s. \end{aligned}$$

Since $1 - \exp(-x) \geq x/2$ for all $x \in [0, 1]$, for ϵ small enough so that $\epsilon/(2b\sqrt{s}) < 1$, we conclude that for $(\tau, \gamma) \in \mathcal{B}$, the integrand in (4.7) can be bounded below as follows:

$$\begin{aligned} & \mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon \mid \tau, \gamma) \\ & \geq (1 - s/p)^{p-s} \exp \left\{ -s \log 2 - \sum_{j \in S_0} \frac{|\theta_{0j}|}{a} + s \log \frac{\epsilon}{4b\sqrt{s}} \right\} \\ (4.8) \quad & \geq e^{-Cs} \exp \left\{ -s \log 2 - \sum_{j \in S_0} \frac{|\theta_{0j}|}{a} + s \log \frac{\epsilon}{4b\sqrt{s}} \right\}, \end{aligned}$$

where the last inequality uses $(1 - x)^{1/x} \geq 1/(2e)$ for $0 \leq x \leq 1/2$ and $C = \log(2e)$. It thus remains to obtain a lower bound to

$$\begin{aligned} \mathbb{P}(\mathcal{B}) &= \int_{(\tau, \gamma) \in \mathcal{B}} f_\gamma(d\tilde{\gamma}) f_\tau(d\tau) \\ (4.9) \quad &= \int_{\tau=2s}^{4s} \mathbb{P}(\mathcal{A}_\tau \mid \tau) f_\tau(d\tau). \end{aligned}$$

Now, since $\gamma \sim \text{Dir}(\alpha/p, \dots, \alpha/p)$, recalling the definition of \mathcal{A}_τ from (4.5)

and using (4.6),

$$\begin{aligned}
\mathbb{P}(A_\tau \mid \tau) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/p)^p} \int_{\tilde{\gamma} \in \mathcal{A}_\tau} \left[\prod_{j=1}^{p-1} \gamma_j^{\alpha/p-1} \right] \left(1 - \sum_{j=1}^{p-1} \gamma_j \right)^{\alpha/p-1} d\gamma_1 \dots d\gamma_{p-1} \\
&\geq C_p (1-b)^{\alpha/p-1} \int_{\tilde{\gamma} \in \mathcal{A}_\tau} \left[\prod_{j \in S_1} \gamma_j^{\alpha/p-1} \right] \times \left[\prod_{j \in S_0^c} \gamma_j^{\alpha/p-1} \right] d\gamma_1 \dots d\gamma_{p-1} \\
(4.10) \quad &\geq C_p (1-b)^{\alpha/p-1} \left\{ \frac{\delta}{\log(p/s)} \right\}^{\alpha(p-s)/p} \left\{ \left(\frac{b}{\tau} \right)^{\alpha/p} - \left(\frac{a}{\tau} \right)^{\alpha/p} \right\}^{s-1},
\end{aligned}$$

where

$$\begin{aligned}
(4.11) \quad C_p &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/p)^p} \left(\frac{p}{\alpha} \right)^{p-1} \\
&= \exp\{\log \Gamma(\alpha) + (p-1) \log(p/\alpha) - p \log \Gamma(\alpha/p)\} \\
&\geq \exp\{\log \Gamma(\alpha) - \log \Gamma(\alpha/p)\} \\
(4.12) \quad &\geq \exp\{\log \Gamma(\alpha) - \log(p/\alpha)\}
\end{aligned}$$

with the last two inequalities using $\Gamma(x) \leq 1/x$ for all $x \in (0, 1)$. Moreover, since $b \geq 4a$, we have for $\tau \in [2s, 4s]$,

$$\begin{aligned}
(4.13) \quad &\left\{ \left(\frac{b}{\tau} \right)^{\alpha/p} - \left(\frac{a}{\tau} \right)^{\alpha/p} \right\}^{s-1} \geq \left\{ \left(\frac{b}{4s} \right)^{\alpha/p} - \left(\frac{a}{2s} \right)^{\alpha/p} \right\}^{s-1} \\
&\geq \left(\frac{b}{4s} \right)^{(s-1)\alpha/p} \left[1 - \exp \left\{ -\frac{\alpha}{p} \log(2b/a) \right\} \right].
\end{aligned}$$

Equations (4.12) and (4.13), in conjunction with the fact that $1 - e^{-x} \geq x/2$ for $x \in (0, 1)$ implies that the expression in (4.10), and thus $\mathbb{P}(\mathcal{A}_\tau \mid \tau)$ in (4.9), is bounded below by

$$(4.14) \quad \mathbb{P}(\mathcal{A}_\tau \mid \tau) \geq C \exp \left\{ \frac{\alpha(p-s)}{p} \log \frac{\delta}{\log(p/s)} - \log \frac{p}{\alpha} - \frac{1}{\log(b/2a)} \log \frac{p}{\alpha} \right\}$$

for some constant $C > 0$. Finally, (4.8) and (4.14) substituted into (4.7) gives us

$$\mathbb{P}(\|\theta - \theta_0\|_2 < \epsilon) \geq \mathbb{P}[\tau \in (2s, 4s)] e^{-C \max\{s \log(s/\epsilon), \log p\}}.$$

The proof of Lemma 4.1 is completed upon observing that $\mathbb{P}[\tau \in (2s, 4s)] \geq e^{-C \log p}$ by definition. \square

We would next like to show that the shrinkage prior in (PS) doesn't spread its mass across too many dimensions. A point mass mixture prior allows a high-dimensional vector to collapse onto fewer dimensions. Hence, the implied dimensionality can be naturally studied through appropriate tail bounds for the induced prior on $|\text{supp}(\theta)|$, which is a random variable supported on $\{0, 1, \dots, p\}$. Such bounds on the prior dimensionality lead to better control of the metric entropy and enable construction of sieves, see Castillo and van der Vaart (2012). However, continuous shrinkage priors do not allow exact zeroes in θ and clearly $\mathbb{P}(|\text{supp}(\theta)| = p) = 1$. Recalling the intuition that (PS) shrinks a large subset of the entries in θ close to zero while allowing a few large signals, we devise a generalized definition of the support of a vector as the subset of entries which are larger than a small number δ in magnitude. For any $\delta > 0$, we denote the corresponding subset to be $\text{supp}_\delta(\theta)$, so that

$$\text{supp}_\delta(\theta) = \{j : |\theta_j| > \delta\}.$$

In the following Lemma 4.3, we provide a tail bound for $\text{supp}_\delta(\theta)$ that is crucially used later in Section 6.

LEMMA 4.3. *Let $\epsilon \in (0, 1)$ and $\delta = \epsilon/p$. If θ is drawn according to the prior (PS), then there exists a constant $A > 0$ such that*

$$\mathbb{P}(|\text{supp}_\delta(\theta)| > A \log p) \leq e^{-C \log p}$$

for some constant $C > 0$.

PROOF. Let $s = \log p$. Clearly, for any $A > 0$,

(4.15)

$$\mathbb{P}(|\text{supp}_\delta(\theta)| > As) = \sum_{j=As}^p \int_{\tau=0}^{\infty} \int_{\tilde{\gamma} \in \Delta_0^{p-1}} \mathbb{P}(|\text{supp}_\delta(\theta)| = j \mid \tau, \gamma) f_\gamma(d\tilde{\gamma}) f_\tau(d\tau).$$

Observe that

$$\mathbb{P}(|\text{supp}_\delta(\theta)| = j \mid \tau, \gamma) = \sum_{S: |S|=j} \prod_{j \in S} \pi_j \prod_{j \in S^c} (1 - \pi_j)$$

where $\pi_j = \mathbb{P}(|\theta_j| > \delta \mid \gamma, \tau) = e^{-\delta/(\tau\gamma_j)}$ by the prior specification in (PS).

We can clearly restrict our attention to $\{\tau \geq \tau_0\}$ with

$$(4.16) \quad \tau_0 = \frac{(p-1)\delta}{\log(p/s)},$$

since by definition of f_τ ,

$$(4.17) \quad \mathbb{P}[\tau < \tau_0] \leq \exp\{-C \log p\}.$$

For a fixed $\tau \geq \tau_0$, consider $\mathcal{E}_\tau \subset \Delta_0^{p-1}$ with,

$$\mathcal{E}_\tau = \left\{ \tilde{\gamma} : \gamma_j \in \left[\frac{\delta}{\tau \log(2p/s)}, \frac{\delta}{\tau \log(p/s)} \right], j = 1, \dots, p-1 \right\}.$$

Clearly, \mathcal{E}_τ defines a valid subset of Δ_0^{p-1} for any $\tau \geq \tau_0$ since $\sum_{j=1}^{p-1} \gamma_j \leq (p-1)\delta/\{\tau \log(p/s)\} \leq 1$ by (4.16). Moreover, on \mathcal{E}_τ , $\pi_j \in [s/2p, s/p]$ for all $j = 1, \dots, p-1$ and thus it follows from Lemma 4.2 of [Castillo and van der Vaart \(2012\)](#) that

$$(4.18) \quad \sum_{j=As}^{p-s} \mathbb{P}(|\text{supp}_\delta(\theta)| = j \mid \tau, \gamma) \leq e^{-Cs}$$

since $(1-x)^{1/x} \leq 1/e$ for all $x \in (0, 1)$. The proof of Lemma 4.3 will be completed if we can show that

$$(4.19) \quad \mathbb{P}(\mathcal{E}_\tau^c \mid \tau) \leq e^{-C \log p}$$

for any $\tau \geq \tau_0$. To that end, proceeding along the lines of the calculations in (4.10),

$$(4.20) \quad \begin{aligned} \mathbb{P}(\mathcal{E}_\tau^c \mid \tau) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/p)^p} \int_{\mathcal{E}_\tau^c} \prod_{j=1}^{p-1} \gamma_j^{\alpha/p-1} \left(1 - \sum_{j=1}^{p-1} \gamma_j\right)^{\alpha/p-1} d\tilde{\gamma} \\ &\leq C_p \left\{ 1 - \left(\frac{\delta}{\tau \log(p/s)} \right)^{\alpha/p} + \left(\frac{\delta}{\tau \log(2p/s)} \right)^{\alpha/p} \right\}^{p-1} \end{aligned}$$

$$(4.21) \quad \times \left\{ 1 - \frac{(p-1)\delta}{\tau \log(2p/s)} \right\}^{\alpha/p-1}$$

with C_p as in (4.11). To obtain an upper bound for C_p , we study the function $g(x) = \frac{1}{x} \log \left(\frac{1}{x\Gamma(x)} \right)$ near zero in the following Lemma 4.4; a proof can be found in the Appendix.

LEMMA 4.4. *The function $g(x) = \frac{1}{x} \log \left(\frac{1}{x\Gamma(x)} \right)$ is monotonically decreasing on $(0, 1/2)$ with $\lim_{x \rightarrow 0} g(x) = \gamma_0$, where $\gamma_0 = -\Gamma'(1)$ is the Euler constant.*

Letting $x = \alpha/p$ in Lemma 4.4, $p \log(p/\alpha) - p \log \Gamma(\alpha/p) = \alpha g(x) \leq \alpha \gamma_0$. Hence,

$$\begin{aligned} C(p) &= \exp\{\log \Gamma(\alpha) + (p-1) \log(p/\alpha) - p \log \Gamma(\alpha/p)\} \\ &\leq C \exp\{-\log(p/\alpha)\} \end{aligned}$$

for some constant $C > 0$. Note that Lemma 4.4 is indeed needed, since the usual $\Gamma(x) \geq 1/(2x)$ on $(0, 1)$ would lead to the less stringent bound $C(p) \leq 2^p$.

Now, for $\tau \geq \tau_0$,

$$\begin{aligned} 0 &\leq \left(\frac{\delta}{\tau \log(p/s)}\right)^{\alpha/p} - \left(\frac{\delta}{\tau \log(2p/s)}\right)^{\alpha/p} \\ &= \left(\frac{\delta}{\tau \log(p/s)}\right)^{\alpha(p-1)/p} \left[1 - \left\{\frac{\log(p/s)}{\log(p/s) + \log 2}\right\}^{\alpha/p}\right] \\ &\leq \left(\frac{1}{p-1}\right)^{\alpha(p-1)/p} \leq e^{-\alpha/2 \log(p-1)} < 1, \end{aligned}$$

implying the second term in (4.20) can be bounded by 1. Equation (4.19) is established upon observing that the term in (4.21) can be bounded above by a constant.

Equations (4.17), (4.18) and (4.19) imply that each summand in (4.15) is bounded above by $e^{-C \log p}$. Noting that there are $(p - As)$ such terms, the proof of Lemma 4.3 follows by choosing A suitably large. \square

A final important property of the proposed shrinkage prior is established through the following large deviation result on the l_1 norm of θ :

LEMMA 4.5. *We have $\mathbb{P}[\|\theta\|_1 \geq (\log p)^2] \leq e^{-C \log p}$.*

PROOF. Recall $\theta_j \sim \text{DE}(\tau \gamma_j)$. Let $X_j = \theta_j/(\tau \gamma_j)$, clearly $X_j \sim \text{DE}(1)$. Let $\psi_j = \tau \gamma_j$ and fix $t > 0$. We now use a Bernstein-type tail inequality for sub-exponential random variables (Proposition 5.16 of Vershynin (2010)) to conclude

$$\begin{aligned} \mathbb{P}\left(\sum_{j=1}^p |\theta_j| > t \mid \tau, \gamma\right) &= \mathbb{P}\left(\sum_{j=1}^p |\psi_j X_j| > t\right) \\ &\leq e^{-C \min\left\{\frac{t^2}{\|\psi\|_2^2}, \frac{t}{\|\psi\|_\infty}\right\}} \leq \max\{e^{-Ct^2/\tau^2}, e^{-Ct/\tau}\}. \end{aligned}$$

The last inequality in the above display uses $\|\gamma\|_2 \leq 1$ and the fact that $e^{-c/x}$ is increasing with x . Thus, with $t = (\log p)^2$,

$$\mathbb{P}\{(\|\theta\|_1 > t) \cap (\tau \leq \log p)\} \leq e^{-C \log p}.$$

The proof is completed since $\mathbb{P}(\tau \geq \log p) \leq e^{-C \log p}$. \square

5. Auxiliary results. In this section, we provide a number of auxiliary results that are used to prove the main results in Section 3 and are also of independent interest.

5.1. Some matrix results. We begin with some matrix inequalities that are used throughout.

LEMMA 5.1. *For any two matrices A, B ,*

- (i) $s_{\min}(A) \|B\|_F \leq \|AB\|_F \leq \|A\|_2 \|B\|_F$
- (ii) $s_{\min}(A) \|B\|_2 \leq \|AB\|_2 \leq \|A\|_2 \|B\|_2$
- (iii) $s_{\min}(A) s_{\min}(B) \leq s_{\min}(AB) \leq \|A\|_2 s_{\min}(B)$.

The next lemma comes handy in manipulating the log-likelihood ratio of two multivariate normal densities.

LEMMA 5.2. *For $p \times p$ positive definite matrices Σ, Σ' ,*

$$(R1) \quad \text{tr}[(\Sigma' \Sigma^{-1} - I_p)^2] = \left\| \Sigma^{-1/2} \Sigma' \Sigma^{-1/2} - I_p \right\|_F^2.$$

$$(R2) \quad \log |\Sigma' \Sigma^{-1}| - \text{tr}(\Sigma' \Sigma^{-1} - I_p) < 0.$$

PROOF. To prove the identity in (R1), observe that by similarity, $\Sigma' \Sigma^{-1}$ and $\Sigma^{-1/2} \Sigma' \Sigma^{-1/2}$ have the same set of non-zero eigenvalues and thus

$$\text{tr}[(\Sigma' \Sigma^{-1} - I_p)^2] = \text{tr}[(\Sigma^{-1/2} \Sigma' \Sigma^{-1/2} - I_p)^2].$$

The proof is completed upon observing $\text{tr}(A^2) = \|A\|_F^2$ for any symmetric matrix A .

To prove (R2), let $\Sigma' = \Sigma + R$, so that $\Sigma' \Sigma^{-1} - I_p = R \Sigma^{-1}$. Since $\Sigma^{-1/2} \Sigma' \Sigma^{-1/2}$ is positive definite, by the similarity argument in the paragraph above, all eigenvalues of $\Sigma' \Sigma^{-1}$ are positive. Let us denote these eigenvalues by $1 + \theta_j, j = 1, \dots, p$ with $\theta_j > -1$. Thus,

$$\log |\Sigma' \Sigma^{-1}| - \text{tr}(\Sigma' \Sigma^{-1} - I_p) = \sum_{j=1}^p \{\log(1 + \theta_j) - \theta_j\} < 0.$$

\square

The next Lemma is adapted from Lemma 5.36 of [Vershynin \(2010\)](#).

LEMMA 5.3. *For a $p \times k$ matrix B with $p > k$, suppose*

$$\|B^T B - I_k\|_2 \leq \max\{\delta, \delta^2\}$$

for some $\delta > 0$. Then,

$$1 - \delta \leq s_{\min}(B) \leq s_{\max}(B) \leq 1 + \delta.$$

5.2. A large deviation result for quadratic forms. Lemma 5.4 below provides an exponential tail bound for the sample average of symmetric quadratic forms around the population mean.

LEMMA 5.4. *Let $\xi_1, \dots, \xi_n \sim N_p(0, I_p)$ and A be a $p \times p$ symmetric matrix. Define $Q_i = \xi_i^T A \xi_i$. Then, for every $t \geq 0$,*

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n Q_i - \text{tr}(A) \right| \geq t \right] \leq 2 \exp \left[-C \min \left(\frac{nt^2}{K^2 \|A\|_F^2}, \frac{nt}{K \|A\|_2} \right) \right]$$

for some absolute constants $C, K > 0$.

PROOF. Since A is symmetric, all eigenvalues of A are real. Let $A = V D V^T$ be an eigendecomposition of A , with V a $p \times p$ orthogonal matrix and $D = \text{diag}(d_1, \dots, d_p)$ a diagonal matrix of the eigenvalues. Letting $\eta_i = V^T \xi_i$, clearly $\eta_i \sim N_p(0, I_p)$ since V is orthogonal. Thus,

$$\frac{1}{n} \sum_{i=1}^n Q_i - \text{tr}(A) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p d_j (\eta_{ij}^2 - 1).$$

We now use Proposition 5.16 of [Vershynin \(2010\)](#) which provides an exponential tail inequality for centered sub-exponential random variables. The proof is completed by noting that $\eta_{ij}^2 - 1$ is centered sub-exponential and $\|A\|_2 = \max_j |d_j|$; $\|A\|_F^2 = \sum_{j=1}^p d_j^2$. \square

A standard approach ([Ghosal, Ghosh and van der Vaart, 2000](#)) in Bayesian asymptotic theory to establish a posterior contraction rates (say ϵ_n) is to develop exponentially consistent test functions for the true density versus the complement of an ϵ_n ball (in an appropriate norm) around the truth with type I and II error rates of the order $\exp(-n\epsilon_n^2)$. This serves as an asymptotic identifiability criterion where the likelihood can differentiate the true density from ones that are ϵ_n apart. The choice of the distance metric plays

a crucial role in dictating the error rates. Accordingly, the next two subsections are devoted towards developing point null versus point alternative test functions in Frobenius and operator norm.

Specifically, based on n i.i.d. samples y_1, \dots, y_n from $N_{p_n}(0, \Sigma_n)$, consider testing

$$(5.1) \quad H_0 : \Sigma_n = \Sigma_{0n} \quad \text{versus} \quad H_1 : \Sigma_n = \Sigma_{1n}, \quad \|\Sigma_{0n} - \Sigma_{1n}\| \geq j\epsilon_n$$

for some integer $j \geq 1$ with $\|\cdot\|$ the Frobenius or the operator norm. We shall denote probabilities/expectations under Σ_{0n} and Σ_{1n} by $\mathbb{E}_0/\mathbb{P}_0$ and $\mathbb{E}_1/\mathbb{P}_1$ respectively.

5.3. Test function construction in Frobenius norm. As mentioned earlier, we consider $p_n < n$ for the Frobenius norm. We show that,

THEOREM 5.5. *Let $\Sigma_{0n}, \Sigma_{1n} \in \mathcal{C}_{0n}$ with $p_n < n$. Let $\bar{\rho}_{ln}, \underline{\rho}_{ln} \geq 1$ be sequences such that*

$$(5.2) \quad \frac{1}{\underline{\rho}_{ln}} \leq s_{\min}(\Sigma_{ln}) \leq s_{\max}(\Sigma_{ln}) \leq \bar{\rho}_{ln}$$

for $l = 0, 1$. Let $\bar{\epsilon}_n = \epsilon_n/\bar{\rho}_{1n}$. Then, there exists a constant $J \geq 1$ such that, for any $j \geq J$, there exist a sequence of test functions $\phi_{j,n}$ for (5.1) with $\|\cdot\| = \|\cdot\|_F$ such that,

$$(5.3) \quad \mathbb{E}_{\Sigma_{0n}} \phi_{j,n} \leq e^{-Cnj^2\bar{\epsilon}_n^2 t_{1n}^2},$$

$$(5.4) \quad \mathbb{E}_{\Sigma_{1n}} (1 - \phi_{j,n}) \leq e^{-Cnj^2\bar{\epsilon}_n^2 t_{2n}^2}$$

for some constant $C > 0$ and $t_{2n} = 1/(\bar{\rho}_{0n}\underline{\rho}_{0n})$, $t_{1n} = t_{2n}^2 \min(\underline{\rho}_{1n}/\bar{\rho}_{1n}, 1/\sqrt{\underline{\rho}_{1n}})$.

PROOF. Let $A_n = \Sigma_{1n}^{1/2}(\Sigma_{0n}^{-1} - \Sigma_{1n}^{-1})\Sigma_{1n}^{1/2}$, $d_n = \|A_n\|_F$, $Q_i = y_i^T(\Sigma_{0n}^{-1} - \Sigma_{1n}^{-1})y_i$, $\bar{Q}_n = n^{-1} \sum_{i=1}^n Q_i$, and $D_n = \log |\Sigma_{1n}\Sigma_{0n}^{-1}|$. Using standard results for quadratic forms,

$$(5.5) \quad \mathbb{E}_0 \bar{Q}_n = \text{tr}(\mathbf{I}_{p_n} - \Sigma_{0n}\Sigma_{1n}^{-1}), \quad \mathbb{E}_1 \bar{Q}_n = \text{tr}(\Sigma_{1n}\Sigma_{0n}^{-1} - \mathbf{I}_{p_n}).$$

We define our test function in terms of the rejection region as $\phi_{j,n} := 1_{[\bar{Q}_n - D_n \geq -\alpha_n d_n^2]}$, where $0 < \alpha_n < 1$ is to be determined in the sequel.

We begin by establishing the type I error bound in (5.3). To that end, we first show that one can find $\beta_n \in (0, 1)$ with $\mathbb{E}_0 \bar{Q}_n - D_n \leq -\beta_n d_n^2$. Clearly,

$$\mathbb{E}_0 \bar{Q}_n - D_n = \log |\Sigma_{0n}\Sigma_{1n}^{-1}| - \text{tr}(\Sigma_{0n}\Sigma_{1n}^{-1} - \mathbf{I}_{p_n}).$$

Let H_n denote the symmetric positive definite matrix $\Sigma_{1n}^{-1/2}\Sigma_{0n}\Sigma_{1n}^{-1/2}$ with eigenvalues $\psi_l > 0, l = 1, \dots, p_n$. By similarity, $\Sigma_{0n}\Sigma_{1n}^{-1}$ has the same eigenvalues as H_n . Also, $A_n = H_n^{-1} - \mathbf{I}_{p_n}$. Thus,

$$(5.6) \quad \begin{aligned} (D_n - \mathbb{E}_0 \bar{Q}_n) - \beta_n d_n^2 &= \text{tr}(H_n - \mathbf{I}_p) - \log |H_n| - \beta_n \left\| H_n^{-1} - \mathbf{I}_{p_n} \right\|_F^2 \\ &= \sum_{l=1}^{p_n} \left[\psi_l - 1 - \log \psi_l - \beta_n \left(\frac{1}{\psi_l} - 1 \right)^2 \right]. \end{aligned}$$

Now, by Lemma 5.1, $s_{\min}(H_n) = s_{\min}(\Sigma_{1n}^{-1}\Sigma_{0n}) \geq s_{\min}(\Sigma_{1n}^{-1})s_{\min}(\Sigma_{0n}) \geq 1/(\underline{\rho}_{0n}\bar{\rho}_{1n})$. Hence, choosing $\beta_n = 1/(\underline{\rho}_{0n}\bar{\rho}_{1n})^2$, one can ensure that the expression in (5.6) is non-negative. Choosing $\alpha_n = \beta_n/2$, we have

$$(5.7) \quad \begin{aligned} \mathbb{E}_0(\phi_{j,n}) &= \mathbb{P}_0(\bar{Q}_n - D_n \geq -\alpha_n d_n^2) \\ &= \mathbb{P}_0\{\bar{Q}_n - \mathbb{E}_0 \bar{Q}_n \geq -\alpha_n d_n^2 - (\mathbb{E}_0 \bar{Q}_n - D_n)\} \\ &\leq \mathbb{P}_0(\bar{Q}_n - \mathbb{E}_0 \bar{Q}_n \geq \beta_n d_n^2/2). \end{aligned}$$

Letting $\xi_i = \Sigma_{0n}^{-1/2}y_i$, it follows that $Q_i = \xi_i^\top B_n \xi_i$ with $B_n = \mathbf{I}_{p_n} - \Sigma_{0n}^{1/2}\Sigma_{1n}^{-1}\Sigma_{0n}^{1/2}$. Clearly, $\xi_i \sim \mathcal{N}(0, \mathbf{I}_{p_n})$ under H_0 . Using (5.5) and invoking Lemma 5.4, we obtain

$$(5.8) \quad \begin{aligned} &\mathbb{P}_0(\bar{Q}_n - \mathbb{E}_0 \bar{Q}_n \geq \beta_n d_n^2/2) \\ &\leq \mathbb{P}_0 \left[\left| \frac{1}{n} \sum_{i=1}^n \xi_i^\top B_n \xi_i - \text{tr}(B_n) \right| \geq \beta_n d_n^2/2 \right] \\ &\leq \exp \left[-C \min \left\{ \frac{n\beta_n^2 d_n^4}{K^2 \|B_n\|_F^2}, \frac{n\beta_n d_n^2}{K \|B_n\|_2} \right\} \right]. \end{aligned}$$

Now, using Lemma 5.1 & (R1) in Lemma 5.2,

$$(5.9) \quad \begin{aligned} d_n^2 &= \left\| \Sigma_{1n}^{1/2} \Sigma_{0n}^{-1} \Sigma_{1n}^{1/2} - \mathbf{I}_{p_n} \right\|_F^2 \\ &= \text{tr}[(\Sigma_{0n}^{-1} \Sigma_{1n} - \mathbf{I}_{p_n})^2] \\ &= \left\| (\Sigma_{0n}^{-1/2} \Sigma_{1n} \Sigma_{0n}^{-1/2} - \mathbf{I}_{p_n}) \right\|_F^2 \\ &\geq s_{\min}(\Sigma_{0n}^{-1})^2 \left\| \Sigma_{1n} - \Sigma_{0n} \right\|_F^2 \\ &= \frac{\left\| \Sigma_{1n} - \Sigma_{0n} \right\|_F^2}{\left\| \Sigma_{0n} \right\|_2^2} \geq \frac{\left\| \Sigma_{1n} - \Sigma_{0n} \right\|_F^2}{\bar{\rho}_{0n}^2}. \end{aligned}$$

Along similar lines,

$$(5.10) \quad \begin{aligned} \|B_n\|_F^2 &\leq s_{\max}(\Sigma_{1n}^{-1})^2 \|\Sigma_{1n} - \Sigma_{0n}\|_F^2 \\ &= \frac{\|\Sigma_{1n} - \Sigma_{0n}\|_F^2}{s_{\min}(\Sigma_{1n})^2} \leq \frac{\|\Sigma_{1n} - \Sigma_{0n}\|_F^2}{\underline{\rho}_{1n}^2}. \end{aligned}$$

Also, by Lemma 5.1,

$$(5.11) \quad \|B_n\|_2 \leq 1 + s_{\max}(\Sigma_{1n}^{-1})s_{\max}(\Sigma_{0n}) \leq 2\bar{\rho}_{0n}\underline{\rho}_{1n}.$$

Thus, from (5.9) - (5.11),

$$(5.12) \quad \frac{\beta_n^2 d_n^4}{\|B_n\|_F^2} \geq \frac{\|\Sigma_{1n} - \Sigma_{0n}\|_F^2}{\bar{\rho}_{1n}^2} \cdot \frac{\underline{\rho}_{1n}^2}{\bar{\rho}_{1n}^2} \cdot \frac{1}{\bar{\rho}_{0n}^4 \underline{\rho}_{0n}^4},$$

$$(5.13) \quad \frac{\beta_n d_n^2}{\|B_n\|_2} \geq \frac{\|\Sigma_{1n} - \Sigma_{0n}\|_F^2}{\bar{\rho}_{1n}^2} \cdot \frac{1}{\underline{\rho}_{1n}} \cdot \frac{1}{\bar{\rho}_{0n}^2 \underline{\rho}_{0n}^2}.$$

Equation (5.3) clearly follows from substituting the bounds obtained in (5.12) - (5.13) in (5.8).

Now on to the type II error. We have

$$(5.14) \quad \begin{aligned} \mathbb{E}_1(\phi_{j,n}) &= \mathbb{P}_1(\bar{Q}_n - \mathbb{E}_1 \bar{Q}_n + \mathbb{E}_1 \bar{Q}_n - D_n \leq -\alpha_n d_n^2) \\ &\leq \mathbb{P}_1(\bar{Q}_n - \mathbb{E}_1 \bar{Q}_n \leq -\alpha_n d_n^2), \end{aligned}$$

where the last inequality uses $\mathbb{E}_1 \bar{Q}_n - D_n > 0$, which is immediate from (R2) in Lemma 5.2. Letting $\xi_i = \Sigma_1^{-1/2} y_i$, proceeding as before and invoking Lemma 5.4 once again, we can upper-bound the expression in (5.14) by

$$(5.15) \quad \exp \left[-C \min \left\{ \frac{n\alpha_n^2 d_n^4}{K^2 \|A_n\|_F^2}, \frac{n\alpha_n d_n^2}{K \|A_n\|_2} \right\} \right] = \exp \left[-C \min \left\{ \frac{n\alpha_n^2 d_n^2}{K^2}, \frac{n\alpha_n d_n^2}{K \|A_n\|_2} \right\} \right],$$

since $d_n = \|A_n\|_F$. Also, by Lemma 5.1,

$$(5.16) \quad \|A_n\|_2 \leq 1 + s_{\max}(\Sigma_{0n}^{-1})s_{\max}(\Sigma_{1n}) \leq 2\bar{\rho}_{1n}\underline{\rho}_{0n}.$$

Thus, up to constants,

$$(5.17) \quad \alpha_n^2 d_n^2 \geq \frac{\|\Sigma_{1n} - \Sigma_{0n}\|_F^2}{\bar{\rho}_{1n}^2} \cdot \frac{1}{\bar{\rho}_{0n}^2 \underline{\rho}_{0n}^2},$$

$$(5.18) \quad \frac{\alpha_n d_n^2}{\|A_n\|_2} \geq \frac{\|\Sigma_{1n} - \Sigma_{0n}\|_F^2}{\bar{\rho}_{1n}^2} \cdot \frac{1}{\bar{\rho}_{0n}^2 \underline{\rho}_{0n}^2}.$$

As before, (5.4) follows from substituting the bounds obtained in (5.17) - (5.18) into (5.15). \square

5.4. Test function construction in operator norm. We now construct a novel test function for large covariance matrices ($p_n \gg n$) that admit a factor decomposition (1.2) to separate points in \mathcal{C}_n in the operator norm. We were unable to use the likelihood ratio test to attain the desired error rates for the operator norm. It seems difficult to exploit the inbuilt parsimony. We instead use a projection technique to design our tests.

THEOREM 5.6. *Let $\Sigma_{0n}, \Sigma_{1n} \in \mathcal{C}_{0n}$ with $\Sigma_{\ell n} = \Lambda_{\ell n} \Lambda_{\ell n}^\top + \sigma_{\ell n}^2 \mathbf{I}_{p_n}$ for $\ell = 0, 1$. Assume that Λ_{0n} satisfies **(A3)** in Assumption 3.1. Let $\bar{\epsilon}_n = \sqrt{\log p_n/n}$ and $\epsilon_n = \sqrt{(\log p_n)^3/n}$. Then, there exists a positive integer $J \geq 1$, such that for any $j \geq J$, one can construct a sequence of test functions $\phi_{j,n}$ for (5.1) with $\|\cdot\| = \|\cdot\|_2$ such that,*

$$(5.19) \quad \mathbb{E}_{\Sigma_{0n}} \phi_{j,n} \leq e^{-Cnj\bar{\epsilon}_n^2},$$

$$(5.20) \quad \mathbb{E}_{\Sigma_{1n}} (1 - \phi_{j,n}) \leq e^{-Cnj\bar{\epsilon}_n^2}$$

for some constant $C > 0$. Moreover, if ϵ_n is changed to $v_n \sqrt{(\log p_n)^3/n}$ for some increasing sequence v_n , the conclusion of the theorem remains valid with $\bar{\epsilon}_n$ modified to $\sqrt{v_n} \sqrt{\log p_n/n}$.

PROOF. Let $x_i = (1/c_n) \Lambda_{0n}^\top y_i$ and $z_i = \Lambda_{0n} x_i$ for $i = 1, \dots, n$, so that $x_i \in \mathbb{R}^{k_n}$ and $z_i \in \mathbb{R}^{p_n}$. Denote

$$\hat{\Sigma}_y = \frac{1}{n} \sum_{i=1}^n y_i y_i^\top, \quad \hat{\Sigma}_x = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top, \quad \hat{\Sigma}_z = \frac{1}{n} \sum_{i=1}^n z_i z_i^\top.$$

Clearly, $\hat{\Sigma}_z = \Lambda_{0n} \hat{\Sigma}_x \Lambda_{0n}^\top$ and $\hat{\Sigma}_x = (1/c_n^2) \Lambda_{0n}^\top \hat{\Sigma}_y \Lambda_{0n}$. With these definitions, letting $\underline{\epsilon}_n = \sqrt{(\log p_n)^2/n}$, we define our test function to be

$$\phi_{j,n} = 1_{\{\|\hat{\Sigma}_z - \Sigma_{0n}\|_2 > j\underline{\epsilon}_n/2\}}.$$

It is known that (see Bickel and Levina (2008a,b); Fan, Fan and Lv (2008); Johnstone (2001); Muirhead) the sample covariance matrix $\hat{\Sigma}_y$ does not have the desired concentration around the population mean in operator norm when $p_n > n$. To circumvent this difficulty, we exploit the near low-rank structure in the truth and replace $\hat{\Sigma}_y$ by $\hat{\Sigma}_z$. Our guiding intuition is that the lower-dimensional $\hat{\Sigma}_x$ should concentrate appropriately around its mean (in operator norm) and we hope to carry through the same concentration to $\hat{\Sigma}_z$, since $\Lambda_{0n}/\sqrt{c_n}$ behaves like an approximate isometry under **(A3)** in Assumption 3.1 (see also Lemma 5.3).

We first show that, there exists a positive integer $J^* \geq 1$ such that for $j \geq J^*$,

$$(5.21) \quad \mathbb{P}_0 \left[\left\| \hat{\Sigma}_z - \Sigma_{0n} \right\|_2 > \frac{j\epsilon_n}{2} \right] \leq e^{-Cnj\epsilon_n^2}.$$

Indeed, we have

$$\begin{aligned} & \left\| \hat{\Sigma}_z - \Sigma_{0n} \right\|_2 \\ &= \sup_{v \in \mathcal{S}^{p-1}} \left| v^T \Lambda_{0n} \hat{\Sigma}_x \Lambda_{0n}^T v - v^T \Lambda_{0n} \Lambda_{0n}^T v - \sigma_{0n}^2 \right| \\ &\leq \sup_{v \in \mathcal{S}^{p-1}} \left| v^T \Lambda_{0n} \hat{\Sigma}_x \Lambda_{0n}^T v - v^T \Lambda_{0n} \Lambda_{0n}^T v - \frac{\sigma_{0n}^2}{c_n} v^T \Lambda_{0n} \Lambda_{0n}^T v \right| \\ &+ \sup_{v \in \mathcal{S}^{p-1}} \left| \sigma_{0n}^2 v^T \left[\frac{1}{c_n} \Lambda_{0n} \Lambda_{0n}^T - \mathbf{I}_{p_n} \right] v \right| \\ &\leq \sup_{w \in \mathbb{R}^k: \|w\|_2 \leq \|\Lambda_{0n}\|_2} \left| w^T \hat{\Sigma}_x w - w^T w - w^T \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} w \right| + \sigma_{0n}^2 \left\| \frac{1}{c_n} \Lambda_{0n} \Lambda_{0n}^T - \mathbf{I}_{p_n} \right\|_2 \\ (5.22) \quad &\leq \|\Lambda_{0n}\|_2 \left\| \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2 + \sigma_{0n}^2 \left\| \frac{1}{c_n} \Lambda_{0n} \Lambda_{0n}^T - \mathbf{I}_{p_n} \right\|_2. \end{aligned}$$

Since $\epsilon_n = \sqrt{(\log p_n)^2/n}$, by **(A3)** and **(A4)**, the second term in (5.22) can be bounded above by $j\epsilon_n/4$ by choosing j larger than some constant J_1 . Thus,

$$(5.23) \quad \mathbb{P}_0 \left[\left\| \hat{\Sigma}_z - \Sigma_{0n} \right\|_2 > \frac{j\epsilon_n}{2} \right] \leq \mathbb{P}_0 \left[\|\Lambda_{0n}\|_2 \left\| \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2 > \frac{j\epsilon_n}{4} \right].$$

Now,

$$\mathbb{E}_0 \hat{\Sigma}_x = \frac{1}{c_n^2} \Lambda_{0n}^T [\Lambda_{0n} \Lambda_{0n}^T + \sigma_{0n}^2 \mathbf{I}_{p_n}] \Lambda_{0n} = \left(\frac{1}{c_n} \Lambda_{0n}^T \Lambda_{0n} \right)^2 + \frac{\sigma_{0n}^2}{c_n} \frac{1}{c_n} \Lambda_{0n}^T \Lambda_{0n}.$$

Hence,

$$\left\| \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2$$

(5.24)

$$\leq \left\| \hat{\Sigma}_x - \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 + \left\| \mathbb{E}_0 \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2$$

(5.25)

$$\leq \left\| \hat{\Sigma}_x - \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 + \left\| \left(\frac{1}{c_n} \Lambda_{0n}^\top \Lambda_{0n} \right)^2 - \mathbf{I}_{k_n} \right\|_2 + \frac{\sigma_{0n}^2}{c_n} \left\| \left(\frac{1}{c_n} \Lambda_{0n}^\top \Lambda_{0n} - \mathbf{I}_{k_n} \right) \right\|_2.$$

To tackle the second term of (5.25), simply note that $\|A - \mathbf{I}_{k_n}\|_2 < \delta$ for A symmetric and some $\delta \in (0, 1)$ implies that $\|A^2 - \mathbf{I}_{k_n}\|_2 \leq 3\delta$. Using this observation and invoking **(A3)** and **(A4)**, we obtain

$$\begin{aligned} & \left\| \left(\frac{1}{c_n} \Lambda_{0n}^\top \Lambda_{0n} \right)^2 - \mathbf{I}_{k_n} \right\|_2 + \frac{\sigma_{0n}^2}{c_n} \left\| \left(\frac{1}{c_n} \Lambda_{0n}^\top \Lambda_{0n} - \mathbf{I}_{k_n} \right) \right\|_2 \\ & \leq C \left(3 + \frac{\sigma_{0n}^2}{c_n} \right) \frac{\sqrt{k_n}}{\sqrt{p_n}}, \end{aligned}$$

for some global constant C . Since $\|\Lambda_{0n}\|_2 \leq 2\sqrt{c_n}$ by (A3), the second term in (5.24) multiplied by $\|\Lambda_{0n}\|_2$ can be thus made smaller than $j\epsilon_n/8$ by choosing j larger than some J_2 . Hence, continuing from (5.23),

$$\mathbb{P}_0 \left[\|\Lambda_{0n}\|_2 \left\| \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2 > \frac{j\epsilon_n}{4} \right] \leq \mathbb{P}_0 \left[\|\Lambda_{0n}\|_2 \left\| \hat{\Sigma}_x - \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 > \frac{j\epsilon_n}{8} \right].$$

By a modification to Theorem 5.39 of Vershynin (2010) (see Remark 5.40), we obtain that for every $t > 0$,

$$(5.26) \quad \mathbb{P}_0 \left[\left\| \hat{\Sigma}_x - \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 > \max\{\delta, \delta^2\} \left\| \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 \right] \leq e^{-C't^2}$$

for $\delta = C\sqrt{k_n/n} + t/\sqrt{n}$ and some global constants $C', C > 0$. Choosing $t = C\sqrt{j \log p_n}$ and using $k_n = O(1)$, we get the desired bound (5.19) if we can show that

$$(5.27) \quad \frac{j\epsilon_n}{8} > C \|\Lambda_{0n}\|_2 \left\| \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 \max \left[\sqrt{\frac{j \log p_n}{n}}, \frac{j \log p_n}{n} \right].$$

Indeed, (5.19) holds for $j > 1$, since $\|\Lambda_{0n}\|_2 \leq 2\sqrt{c_n}$, $c_n = O(\log p_n)$, $\left\| \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 = O(1)$, $\sqrt{j} < j$ and $\bar{\epsilon}_n = \sqrt{\log p_n/n} \in (0, 1)$ so that $\bar{\epsilon}_n^2 < \bar{\epsilon}_n$. The claim in (5.21) will thus follow by choosing $J^* = \max\{J_1, J_2\}$.

We next show that, there exists a constant J^{**} such that for $j \geq J^{**}$,

$$(5.28) \quad \mathbb{P}_1 \left[\left\| \hat{\Sigma}_z - \Sigma_{0n} \right\|_2 \leq \frac{j\epsilon_n}{2} \right] \leq e^{-Cnj^2\epsilon_n^2}.$$

Proceeding as in (5.22), we obtain

$$\left\| \hat{\Sigma}_z - \Sigma_{0n} \right\|_2 \geq \|\Lambda_{0n}\|_2 \left\| \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2 - \sigma_{0n}^2 \left\| \frac{1}{c_n} \Lambda_{0n} \Lambda_{0n}^\top - \mathbf{I}_{p_n} \right\|_2.$$

As before, the second term in the right hand side of the above equation can be bounded above by $j\epsilon_n/32$ by choosing j larger than some constant J_3 . Thus,

$$(5.29) \quad \begin{aligned} \mathbb{P}_1 \left[\left\| \hat{\Sigma}_z - \Sigma_{0n} \right\|_2 \leq \frac{j\epsilon_n}{2} \right] &\leq \mathbb{P}_1 \left[\|\Lambda_{0n}\|_2 \left\| \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2 \leq \frac{17j\epsilon_n}{32} \right] \\ &\leq \mathbb{P}_1 \left[\|\Lambda_{0n}\|_2 \left\{ \left\| \mathbb{E}_1 \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2 - \left\| \hat{\Sigma}_x - \mathbb{E}_1 \hat{\Sigma}_x \right\|_2 \right\} \leq \frac{17j\epsilon_n}{32} \right]. \end{aligned}$$

By **(A3)** and Lemma 5.3, both $\|\Lambda_{0n}/\sqrt{c_n}\|_2$ and $s_{\min}(\Lambda_{0n}^\top \Lambda_{0n}/c_n)$ can be bounded below by $3/4$. Further, by (5.24), $\|\Lambda_{0n}\|_2 \left\| \mathbb{E}_0 \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2$ can be bounded above by $j\epsilon_n/64$ by choosing j larger than some constant J_4 . Thus, for $j \geq J^{**} := \max\{J_3, J_4\}$,

$$\begin{aligned} \|\Lambda_{0n}\|_2 \left\| \mathbb{E}_1 \hat{\Sigma}_x - \mathbf{I}_{k_n} - \frac{\sigma_{0n}^2}{c_n} \mathbf{I}_{k_n} \right\|_2 &\geq \|\Lambda_{0n}\|_2 \left\| \mathbb{E}_1 \hat{\Sigma}_x - \mathbb{E}_0 \hat{\Sigma}_x \right\|_2 - \frac{j\epsilon_n}{64} \\ &= \|\Lambda_{0n}\|_2 \left\| \frac{1}{c_n^2} \Lambda_{0n}^\top (\Sigma_{1n} - \Sigma_{0n}) \Lambda_{0n} \right\|_2 - \frac{j\epsilon_n}{64} = \left\| \frac{\Lambda_{0n}}{c_n} \right\|_2 \left\| \frac{1}{c_n} \Lambda_{0n}^\top (\Sigma_{1n} - \Sigma_{0n}) \Lambda_{0n} \right\|_2 - \frac{j\epsilon_n}{64} \\ &\geq \frac{3}{4\sqrt{c_n}} \left\| \frac{1}{c_n} \Lambda_{0n} \Lambda_{0n}^\top (\Sigma_{1n} - \Sigma_{0n}) \right\|_2 - \frac{j\epsilon_n}{64} \geq \frac{3}{4\sqrt{c_n}} \|\Sigma_{1n} - \Sigma_{0n}\|_2 s_{\min} \left(\frac{\Lambda_{0n} \Lambda_{0n}^\top}{c_n} \right) - \frac{j\epsilon_n}{64} \\ &\geq \frac{35j\epsilon_n}{64}, \end{aligned}$$

where the penultimate inequality uses (ii) in Lemma 5.1 and the fact that $\epsilon_n/\sqrt{c_n} \geq \epsilon_n$. Hence, the quantity in (5.29) can be bounded above by

$$\mathbb{P}_1 \left[\|\Lambda_{0n}\|_2 \left\| \hat{\Sigma}_x - \mathbb{E}_1 \hat{\Sigma}_x \right\|_2 \geq \frac{j\epsilon_n}{64} \right],$$

whose treatment follows in a similar fashion as (5.26).

When $\epsilon_n = v_n \sqrt{(\log p_n)^3/n}$ for some increasing sequence v_n , change ϵ_n to $v_n \sqrt{(\log p_n)^2/n}$ in the definition of the test function. The rest of the proof goes through similarly, with the modification that we now have to choose $t = C \sqrt{v_n} \sqrt{j \log p_n}$ in the display following (5.26), leading to $\bar{\epsilon}_n = \sqrt{v_n} \sqrt{\log p_n/n}$. \square

6. Proof of the main results. We now proceed to prove the results stated in Section 3. We prove the rate theorem for the operator norm with shrinkage prior (Theorem 3.6) in details and sketch the argument for the point mass mixture prior (Theorem 3.5). For Theorem 3.3 concerning the Frobenius norm, again only a sketch is provided.

6.1. *Proof of Theorem 3.6.* For $\epsilon_n = \sqrt{(\log p_n)^5/n}$ and some constant $M > 0$, define the set

$$U_n = \{\Sigma_n : \|\Sigma_n - \Sigma_{0n}\|_2 \leq M\epsilon_n\}.$$

The posterior probability assigned to the complement of U_n is given by

$$(6.1) \quad \Pi_n(U_n^c \mid \mathbf{y}^{(n)}) = \frac{\int_{U_n^c} \prod_{i=1}^n \frac{f_{\Sigma_n}(y_i)}{f_{\Sigma_{0n}}(y_i)} d\Pi_n(\Sigma_n)}{\int \prod_{i=1}^n \frac{f_{\Sigma_n}(y_i)}{f_{\Sigma_{0n}}(y_i)} d\Pi_n(\Sigma_n)} \equiv \frac{\mathcal{N}_n}{\mathcal{D}_n},$$

where f_{Σ_n} denotes a p_n -dimensional $N(0, \Sigma_n)$ distribution. Here \mathcal{N}_n and \mathcal{D}_n denote the numerator and denominator of the fraction in (6.1)

Let $\sigma(y_1, \dots, y_n)$ denote the σ -field generated by y_1, \dots, y_n . We first show that we can lower-bound \mathcal{D}_n on an event $A_n \in \sigma(y_1, \dots, y_n)$ with large probability under $f_{\Sigma_{0n}}$ in Lemma 6.1; the proof can be found in the Appendix.

LEMMA 6.1. *Let Σ_{0n} satisfy Assumption 3.4. Let δ_n be a sequence satisfying $\delta_n/s_{\min}(\Sigma_{0n}) \rightarrow 0$ and $n\delta_n^2/s_{\min}(\Sigma_{0n})^2 \rightarrow \infty$. Then, there exists $A_n \in \sigma(y_1, y_2, \dots, y_n)$ with $\mathbb{P}_{\Sigma_{0n}}(A_n) \rightarrow 1$ such that on A_n ,*

$$\mathcal{D}_n \geq e^{-Cn\delta_n^2/s_{\min}(\Sigma_{0n})^2} \Pi_n(\Sigma_n : \|\Sigma_n - \Sigma_{0n}\|_F < \delta_n).$$

By Lemma 6.1, it is enough to show

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\Sigma_{0n}} \left[\Pi_n(\|\Sigma_n - \Sigma_{0n}\|_2 > M\epsilon_n \mid \mathbf{y}^{(n)}) 1_{A_n} \right] = 0$$

to prove Theorem 3.5.

By Assumption 3.1, Σ_{0n} has a low-rank structure $\Lambda_{0n}\Lambda_{0n}^T + \sigma_{0n}^2 \mathbf{I}_{p_n}$ with the true number of factors k_{0n} assumed to be bounded and known in (A4) of Assumption 3.4. Also, recall the model (2.1) is fitted with $k_n = k_{0n}$ factors.

Recall the supp_δ notation from Section 4. Following the convention in Section 2, let $\text{supp}_{\delta'_n}(\Lambda_n)$ denote the set $S \subset \{1, \dots, p_n k_n\}$ corresponding to the entries in Λ_n (vectorized) larger than δ'_n in magnitude. Then, for some $H > 0$ and sequences t_n, δ'_n to be chosen later,

$$\begin{aligned} & \mathbb{E}_{\Sigma_{0n}} \left[\Pi_n(\|\Sigma_n - \Sigma_{0n}\|_2 > M\epsilon_n \mid \mathbf{y}^{(n)}) 1_{A_n} \right] \leq \\ & \mathbb{E}_{\Sigma_{0n}} \left[\mathbb{P}(\|\Sigma_n - \Sigma_{0n}\|_2 > M\epsilon_n, |\text{supp}_{\delta'_n}(\Lambda_n)| \leq H \log p_n, \|\Lambda_n\|_2 \leq t_n, \sigma^2 \leq t_n \mid \mathbf{y}^{(n)}) 1_{A_n} \right] + \\ (6.2) \quad & \mathbb{E}_{\Sigma_{0n}} \Pi_n(|\text{supp}_{\delta'_n}(\Lambda_n)| > H \log p_n \mid \mathbf{y}^{(n)}) + \mathbb{E}_{\Sigma_{0n}}(\|\Lambda_n\|_2 > t_n \mid \mathbf{y}^{(n)}) + \mathbb{E}_{\Sigma_{0n}} \Pi_n(\sigma^2 > t_n \mid \mathbf{y}^{(n)}). \end{aligned}$$

Let $t_n = C(\log p_n)^2$, $\delta_n = C\sqrt{\log p_n/n}$ and $\delta'_n = \delta_n/p_n$. With these choices, we shall first show in Lemma 6.2 and Lemma 6.3 that the posterior probabilities of the sets $\{|\text{supp}_{\delta'_n}(\Lambda_n)| > H \log p_n\}$, $\{\|\Lambda_n\|_2 > t_n\}$ and $\{\sigma^2 > t_n\}$ go to zero, so that we can focus on the set $U_n^* = \{\|\Sigma_n - \Sigma_{0n}\|_2 > M\epsilon_n, \|\Lambda_n\|_2 \leq t_n, \sigma^2 \leq t_n\} \cap \{|\text{supp}_{\delta'_n}(\Lambda_n)| \leq H \log p_n\}$. This will be crucial in reducing the entropy of the model space later on. The proofs for both the Lemmas are provided in the Appendix.

LEMMA 6.2. *Recall $\delta_n = C\sqrt{\log p_n/n}$ and $\delta'_n = \delta_n/p_n$. Then, there exists a constant $H > 0$ such that*

$$(6.3) \quad \lim_{n \rightarrow \infty} \mathbb{E}_{\Sigma_{0n}} [\Pi_n(|\text{supp}_{\delta'_n}(\Lambda_n)| > H \log p_n \mid \mathbf{y}^{(n)}) 1_{A_n}] = 0.$$

LEMMA 6.3. *There exists a constant $C > 0$ such that with $t_n = C(\log p_n)^2$,*

$$(6.4) \quad \lim_{n \rightarrow \infty} \mathbb{E}_{\Sigma_{0n}} [\Pi_n(\|\Lambda_n\|_2 > t_n \mid \mathbf{y}^{(n)}) 1_{A_n}] = 0$$

$$(6.5) \quad \lim_{n \rightarrow \infty} \mathbb{E}_{\Sigma_{0n}} [\Pi_n(\sigma^2 > t_n \mid \mathbf{y}^{(n)}) 1_{A_n}] = 0.$$

We now turn to proving Theorem 3.6. Let $S_0 = \text{supp}(\Lambda_{0n})$. For a set $S \subset \{1, \dots, p_n k_n\}$ with $|S| \leq H \log p_n$, let $B_{j,S,n}$ denote the subset of \mathcal{C}_n :

$$B_{j,S,n} = \{\Sigma_n \in U_n^* : j\epsilon_n < \|\Sigma_n - \Sigma_{0n}\|_2 \leq (j+1)\epsilon_n, \text{supp}_{\delta'_n}(\Lambda_n) = S\}.$$

Then,

$$\begin{aligned}
\mathbb{E}_{\Sigma_{0n}} \mathbb{P}(U_n^* | \mathbf{y}^{(n)}) 1_{A_n} &\leq \sum_{S: |S| \leq H \log p_n} \sum_{j=M}^{\infty} \mathbb{E}_{\Sigma_{0n}} \mathbb{P}[\Sigma_n \in B_{j,S,n} | \mathbf{y}^{(n)}] 1_{A_n} \\
&\leq \sum_{S: |S| \leq H \log p_n} \sum_{j=M}^{\infty} \left[\mathbb{E}_{\Sigma_{0n}} \Phi_{j,S,n} + \mathbb{E}_{\Sigma_{0n}} \left\{ (1 - \Phi_{j,S,n}) \frac{\int_{B_{j,S,n}} \prod_{i=1}^n \frac{f_{\Sigma_n}(y_i)}{f_{\Sigma_{0n}}(y_i)} d\Pi_n(\Sigma_n)}{\mathcal{D}_n} 1_{A_n} \right\} \right] \\
(6.6) \quad &\leq \sum_{S: |S| \leq H \log p_n} \sum_{j=M}^{\infty} \left[\mathbb{E}_{\Sigma_{0n}} \Phi_{j,S,n} + \beta_{j,S,n} \sup_{\Sigma_n \in B_{j,S,n}} \mathbb{E}_{\Sigma_n} (1 - \Phi_{j,S,n}) \right],
\end{aligned}$$

where $\Phi_{j,S,n}$ is a test function for

$$(6.7) \quad H_0 : \Sigma_n = \Sigma_{0n} \quad \text{versus} \quad H_1 : \Sigma_n \in B_{j,S,n}$$

whose construction is provided below and

$$(6.8) \quad \beta_{j,S,n} = \frac{\Pi_n(B_{j,S,n})}{e^{-n\delta_n^2/s_{\min}(\Sigma_{0n})^2} \mathbb{P}(\|\Sigma_n - \Sigma_{0n}\|_F < \delta_n)}.$$

To construct the test function $\Phi_{j,S,n}$, we break up $B_{j,S,n}$ into balls and obtain local tests for Σ_{0n} versus the centers of each of the balls using Theorem 5.6. Since we have already conditioned on $|\text{supp}_{\delta'_n}(\Lambda_n)| \leq H \log p_n$, the number of such balls can be controlled and $\Phi_{j,S,n}$ is obtained as the maximum of the local tests.

Let $\Sigma_{n,l}$ for $l \in I_{j,S,n}$ be a $j\epsilon_n/2$ -net of $B_{j,S,n}$ in operator norm and for each l , define $E_{j,l} = \{\Sigma_n \in B_{j,S,n} : \|\Sigma_n - \Sigma_{n,l}\|_2 \leq j\epsilon_n/2\}$. By definition,

$$B_{j,S,n} \subset \cup_{l \in I_{j,S,n}} E_{n,l}.$$

Let $\phi_{j,S,n,l}$ denote the point versus point test developed in Theorem 5.6 for $\Sigma_n = \Sigma_{0n}$ versus $\Sigma_n = \Sigma_{n,l}$ with the sequence $v_n = \log p_n$, so that $\bar{\epsilon}_n = \sqrt{(\log p_n)^2/n}$ in (5.19) and (5.20). Clearly, $\phi_{j,S,n,l}$ used as a test for Σ_{0n} versus $\Sigma_n \in E_{j,S,n,l}$ retains the same type I and II error rates. Letting $\Phi_{j,S,n} = \max_{l \in I_{j,S,n}} \phi_{j,S,n,l}$, one clearly has from Theorem 5.6,

$$\begin{aligned}
\mathbb{E}_{\Sigma_{0n}}(\Phi_{j,S,n}) &\leq |I_{j,S,n}| e^{-Cnj\bar{\epsilon}_n^2}, \\
\sup_{\Sigma_n \in B_{j,S,n}} \mathbb{E}_{\Sigma_n}(1 - \Phi_{j,S,n}) &\leq e^{-Cnj\bar{\epsilon}_n^2}.
\end{aligned}$$

To estimate $|I_{j,S,n}|$, i.e., the covering number of $B_{j,S,n}$ in operator norm, we first embed $B_{j,S,n}$ inside a bigger set $\tilde{B}_{j,S,n}$ in Lemma 6.4. As we shall see, it is easier to estimate the covering number of $\tilde{B}_{j,S,n}$. For notational convenience, we use $P_S(\theta)$ below to denote θ_S defined in Section 2.

LEMMA 6.4. Recall the sequence t_n from Lemma 6.3. Then,

$$B_{j,S,n} \subset \tilde{B}_{j,S,n} := \{\Sigma_n : \Sigma_n = \Lambda_n \Lambda_n^\top + \sigma_n^2 \mathbf{I}_{p_n}, \Lambda_n \in B_{j,S,n}^{(\Lambda)}, \sigma_n^2 \leq t_n\}$$

where $B_{j,S,n}^{(\Lambda)} = \{\Lambda_n : \text{supp}_{\delta'_n}(\Lambda_n) = S, \|\Lambda_n\|_F \leq Ct_n\}$.

PROOF. The proof simply follows from the fact that $\|\Lambda_n\|_F \leq \sqrt{k_n} \|\Lambda_n\|_2 \leq \sqrt{k_n} t_n \leq Ct_n$ since $k_n = O(1)$ by **(A0)**. \square

We now proceed to explicitly construct a $j\epsilon_n/2$ -net for $\tilde{B}_{j,S,n}$. Let $\xi_n = j\epsilon_n/(4t_n)$. Let $\{\Lambda_l\}_{l=1}^L$ be a ξ_n -net of $B_{j,S,n}^{(\Lambda)}$. Also, let $\{\sigma_r^2\}_{r=1}^R$ be a $j\epsilon_n/4$ -net of $[0, t_n]$. We show below that $\{\Lambda_l \Lambda_l^\top + \sigma_r^2\}_{l,r}$ form a $j\epsilon_n/2$ -net of $\tilde{B}_{j,S,n}$ in operator norm.

Let $\tilde{\Sigma} = \tilde{\Lambda} \tilde{\Lambda}^\top + \tilde{\sigma}^2 \mathbf{I}$ be in $\tilde{B}_{j,S,n}$. Find Λ_l and σ_r^2 from the respective nets so that $\|\Lambda_l - \tilde{\Lambda}\|_F \leq \xi_n$ and $|\sigma_r^2 - \tilde{\sigma}^2| \leq j\epsilon_n/4$. Let $\Sigma = \Lambda_l \Lambda_l^\top + \sigma_r^2$. Then,

$$\begin{aligned} \|\Sigma - \tilde{\Sigma}\|_2 &\leq j\epsilon_n/4 + \|\Lambda_l \Lambda_l^\top - \tilde{\Lambda} \tilde{\Lambda}^\top\| \\ &\leq j\epsilon_n/4 + [\|\Lambda_l\|_2 + \|\tilde{\Lambda}\|_2] \xi_n \leq j\epsilon_n/4 + 2t_n j\epsilon_n/(4t_n) = j\epsilon_n/2. \end{aligned}$$

We have thus proved our claim and hence the $j\epsilon_n/2$ -covering number of $B_{j,S,n}$ is bounded by $L \times R$. Note that the control on $\|\Lambda_n\|_2$ in $B_{j,S,n}$ is crucially used in the above display.

Clearly $R \leq Ct_n/(j\epsilon_n)$. With $s = |S|$, let $\{\theta_l\}_{l=1}^L$ be a $\xi_n/2$ -net of the Euclidean sphere in \mathbb{R}^s of radius Ct_n . By Lemma 5.2 of Vershynin (2010), the cardinality of such a net $L \leq (1 + Ct_n/\xi_n)^s$. We now exhibit a ξ_n -net $\{\Lambda_l\}_{l=1}^L$ to $B_{j,S,n}^{(\Lambda)}$ in Frobenius norm (or equivalently the Euclidean norm after vectorizing) as follows. Set $P_S(\Lambda_l) = \theta_l$ and $P_{S^c}(\Lambda_l) = \mathbf{0}$. Let $\Lambda \in B_{j,S,n}^{(\Lambda)}$ and $\theta = P_S(\Lambda)$. There exists θ_l such that $\|\theta - \theta_l\|_2 \leq \xi_n/2$. Also, since $\text{supp}_{\delta'_n}(\Lambda) = S$, $\|P_{S^c}(\Lambda)\|_2 \leq \delta_n$. By choosing j larger than some constant J , we can make $\xi_n \geq 2\delta_n$. Hence $\|\Lambda_l - \Lambda\|_F \leq \xi_n$.

Thus, finally

$$(6.9) \quad \mathbb{E}_{\Sigma_{0n}}(\Phi_{j,S,n}) \leq e^{Cs \log n} e^{-C_1 n j \bar{\epsilon}_n^2},$$

$$(6.10) \quad \sup_{\Sigma_n \in B_{j,S,n}} \mathbb{E}_{\Sigma_n}(1 - \Phi_{j,S,n}) \leq e^{-C_2 n j \bar{\epsilon}_n^2}.$$

We next proceed to upper-bound $\beta_{j,S,n}$ from (6.8). To that end, we first lower-bound $\mathbb{P}(\|\Sigma_n - \Sigma_{0n}\|_F \leq \delta_n)$ in the following Lemma 6.5.

LEMMA 6.5. *If Σ_{0n} satisfies Assumption 3.4, the prior on Σ_n is as in Theorem 3.6 and $\delta_n = \sqrt{\log p_n/n}$, then*

$$\mathbb{P}(\|\Sigma_n - \Sigma_{0n}\|_F \leq \delta_n) \geq e^{-C \log p_n}.$$

PROOF. It follows that

$$\|\Lambda_n - \Lambda_{0n}\|_F < \delta_n/(4\sqrt{c_n}), |\sigma_n^2 - \sigma_{0n}^2| < \delta_n/(2\sqrt{p_n}) \implies \|\Sigma_n - \Sigma_{0n}\|_F < \delta_n,$$

since, invoking **(A3)** and by Lemma 5.1,

$$\|\Lambda_n \Lambda_n^\top - \Lambda_{0n} \Lambda_{0n}^\top\|_F \leq \left\{ 2 \|\Lambda_{0n}\|_2 + \frac{\delta_n}{4\sqrt{c_n}} \right\} \frac{\delta_n}{4\sqrt{c_n}} \leq \delta_n/2.$$

Now, $\mathbb{P}\{|\sigma_n^2 - \sigma_{0n}^2| < \delta_n/(2\sqrt{p_n})\} \geq \exp(-\sigma_{0n}^2)[1 - \exp\{-\delta_n/(2\sqrt{p_n})\}]$. Using $1 - e^{-x} \geq x/2$ for $x \in (0, 1)$, this term can be bounded below by $e^{-C \log p_n}$.

By **(A3)**, $\|\Lambda_{0n}\|_2 \leq 2\sqrt{c_n}$, implying $\|\Lambda_{0n}\|_F \leq 2\sqrt{c_n k_n}$. Letting $\text{vec}(\Lambda_{0n})$ denote Λ_{0n} vectorized, it follows by **(A2)** and the Cauchy–Schwartz inequality that $\|\text{vec}(\Lambda_{0n})\|_1 \leq C \log p_n$. Hence we can invoke Lemma 4.1 to conclude that $\mathbb{P}\{\|\Lambda_n - \Lambda_{0n}\|_F \leq \delta_n/(4\sqrt{c_n})\} \geq e^{-C \log p_n}$.

□

Using Lemma 6.5 and **(A4)**, $\beta_{j,S,n} \leq e^{-C \log p_n}$. Substituting the error bounds obtained in (6.9) & (6.10) in (6.6), we can bound the expression in (6.6) by

$$(6.11) \quad \sum_{s=0}^{H \log p_n} \binom{p_n}{s} \left[\sum_{j=M}^{\infty} e^{Cs \log n} e^{-C_1 j (\log p_n)^2} + \beta_{j,S,n} e^{-C_2 j (\log p_n)^2} \right].$$

Noting that $\max_{\{0 \leq l \leq H \log p_n\}} \binom{p_n}{l} \leq \exp\{C(\log p_n)^2\}$ and substituting the upper bound to $\beta_{j,S,n}$ obtained above in (6.11), (6.11) goes to 0 for large enough $M > 0$. This finishes the proof of Theorem 3.6. □

6.2. *Proof of Theorem 3.5.* The proof is very similar to the proof of the previous theorem, hence we only sketch a brief outline. Since the point mass mixture priors allow exact zeros in the loadings, we can condition on $\text{supp}(\Lambda_n) = S$ here. By properties of point mass mixture priors shown in Castillo and van der Vaart (2012), analogues of Lemmata 6.2, 6.3 and 6.5 can be obtained to conclude the theorem.

6.3. *Proof of Theorem 3.3*. Without loss of generality assume $k_n = 1$. Let $\xi_{1n} = O(1/\log n)$ and $\xi_{2n} = O(p_n)$, so that

$$\begin{aligned}\mathbb{P}(\sigma_n^2 \leq \xi_{1n}) &= \mathbb{P}(\sigma_n^{-2} > \xi_{1n}^{-1}) = 0 \text{ for all large } n, \\ \mathbb{P}(s_{\max}(\Sigma_n) \geq \xi_{2n}) &\leq \exp\{-\xi_{2n}\}.\end{aligned}$$

The second tail probability above follows from known large deviation results for the largest eigenvalue of i.i.d. $N(0, 1)$ $p_n \times k_n$ random matrices; see Corollary 5.35 of [Vershynin \(2010\)](#). Proceeding as in the proofs of Lemma 6.3 and Lemma 6.2 with $\delta_n = \sqrt{p_n/n}$ we conclude that

$$\mathbb{E}_{\Sigma_{0n}} \mathbb{P}(\sigma_n^2 \leq \xi_{1n} \mid \mathbf{y}^n) \rightarrow 0, \quad \mathbb{E}_{\Sigma_{0n}} \mathbb{P}(s_{\max}(\Sigma_n) \geq \xi_{2n} \mid \mathbf{y}^n) \rightarrow 0.$$

Defining $B_{j,n} = \{\Sigma_n : j\epsilon_n < \|\Sigma_{0n} - \Sigma_n\|_F \leq (j+1)\epsilon_n\}$ and denoting D_n as in Lemma 6.1,

$$\begin{aligned}(6.12) \quad & \mathbb{E}_{\Sigma_{0n}} \mathbb{P}\{\|\Sigma_n - \Sigma_{0n}\|_F > M\epsilon_n, \sigma_n^2 > \xi_{1n}, s_{\max}(\Sigma_n) \leq \xi_{2n} \mid \mathbf{y}^{(n)}\} I_{A_n} \\ & \leq \sum_{j=M}^{\infty} \left[\mathbb{E}_{\Sigma_{0n}} \Phi_{j,n} + \beta_{j,n} \sup_{\Sigma_n \in B_{j,n}} \mathbb{E}_{\Sigma_n} (1 - \Phi_{j,n}) \right],\end{aligned}$$

where

$$(6.13) \quad \beta_{j,n} = \frac{\Pi_n(B_{j,n})}{e^{-n\delta_n^2} \Pi_n(\|\Sigma_n - \Sigma_{0n}\|_F < \delta_n)}$$

and $\Phi_{j,n}$ is a test function for

$$(6.14) \quad H_0 : \Sigma_n = \Sigma_{0n} \quad \text{versus} \quad H_1 : \Sigma_n \in B_{j,n}.$$

Observe that the test function in Theorem 5.5 can also be used for testing

$$(6.15) \quad H_0 : \Sigma_n = \Sigma_{0n} \quad \text{versus} \quad H_1 : \Sigma_{1n} \in E_n$$

with $E_n = \{\Sigma_n : \|\Sigma_n - \Sigma_{1n}\|_F \leq \|\Sigma_{1n} - \Sigma_{0n}\|_F / 2\}$. Taking the maximum of the test functions for Σ_{0n} versus each of the balls of type E_n of radius $j\epsilon_n/2$ covering $B_{j,n}$, we obtain $\Phi_{j,n}$ for testing Σ_{0n} versus $B_{j,n}$. From Theorem 5.5 it follows that

$$\begin{aligned}\mathbb{E}_{\Sigma_{0n}} \Phi_{j,n} &\leq N(j\epsilon_n/2, B_{j,n}, \|\cdot\|_F) e^{-Cnj^2\epsilon_n^2 q_{1n}^2}, \\ \mathbb{E}_{\Sigma_{1n}} (1 - \Phi_{j,n}) &\leq e^{-Cnj^2\epsilon_n^2 q_{2n}^2}.\end{aligned}$$

where

$$q_{2n}^2 = \frac{1}{\bar{\rho}_{1n}^2 \bar{\rho}_{0n}^2 \varrho_{0n}^2} = \frac{1}{\xi_{2n}^2 p_n^2 (\log n)^2}, \quad q_{1n}^2 = \frac{\varrho_{1n}^2}{\bar{\rho}_{1n}^4 \bar{\rho}_{0n}^4 \varrho_{0n}^4} = \frac{1}{\xi_{1n}^2 \xi_{2n}^4 p_n^4 (\log n)^4}.$$

Also, from Lemma 5.2 of [Vershynin \(2010\)](#),

$$(6.16) \quad N(\epsilon_n, B_{j,n}, \|\cdot\|_F) \leq \{1 + 4(j+1)/j\}^{p_n} \leq 9^{p_n}.$$

Defining ν_p as the volume of the p -dimensional Euclidean ball, it follows from Lemma 5.2 in [Castillo and van der Vaart \(2012\)](#) that

$$\begin{aligned} \Pi(B_{j,n}) &\leq \nu_{p_n}\{(j+1)\epsilon_n\}^{p_n} \max \left\{ \prod_j \phi_\sigma(\lambda_j) : \|\Lambda_n - \Lambda_{0n}\| < (j+1)\epsilon_n \right\} \\ &\leq \exp \left\{ \frac{p_n}{2} \log C - \frac{(p_n+1)}{2} \log p_n - p_n + p_n \log\{(j+1)\epsilon_n\} \right\}. \end{aligned}$$

From Lemma [A.2](#) with $\kappa_{2n} = O(\sqrt{p_n})$,

$$(6.17) \quad \beta_{j,n} \leq \Pi(B_{j,n}) e^{n\delta_n^2} \exp\{-Cp_n + p_n + p_n \log(\delta_n/2Cp_n)\}$$

for some constant $C > 0$. From (6.16), it follows that (6.12) is bounded above by

$$(6.18) \quad \sum_{j=M}^{\infty} \left\{ e^{p_n \log 9} e^{-j^2 n q_{1n}^2 \epsilon_n^2} + \beta_{j,n} e^{-j^2 n q_{2n}^2 \epsilon_n^2} \right\}$$

which converges to 0 for large enough M if $\epsilon_n = \sqrt{\frac{p_n^9}{n} \log^3 n}$.

7. Acknowledgement. Dr. Debdeep Pati, Dr. Anirban Bhattacharya and Dr. David B. Dunson is partially supported by the DARPA MSEE program and the grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). Dr. Natesh S. Pillai is partially funded by NSF DMS 1107070. The authors would like to thank Steven Finch for careful proofreading of the paper.

APPENDIX

LEMMA A.1. *Let $a = \log p$ and f_τ denote the $\text{IG}(a, a)$ distribution. If $\tau \sim f_\tau$, then for large p ,*

$$\begin{aligned} \mathbb{P}(\tau > \log p) &\leq e^{-C \log p} \\ \mathbb{P}(\tau \in [2 \log p, 4 \log p]) &\geq e^{-C \log p} \\ \mathbb{P}(\tau < 1/\log p) &\leq e^{-C \log p} \end{aligned}$$

where $C > 0$ denotes a (different) constant in each display.

PROOF. Let $a = \log p$ and assume $\tau \sim \text{IG}(a, a)$. Clearly $X = 1/\tau \sim \text{Gamma}(a, a)$ with a density $f(x) = ce^{-ax}x^{a-1}$ on $(0, \infty)$, where $c = a^a/\Gamma(a)$.

We shall use the following result for the gamma function from [Kruijer, Rousseau and van der Vaart \(2010\)](#): For any $\alpha > 0$,

$$\Gamma(\alpha) = \sqrt{2\pi}e^{-\alpha}\alpha^{\alpha-1/2}e^{\theta(\alpha)}$$

where $0 < \theta(\alpha) < 1/(12\alpha)$. Clearly for large α ,

$$C_1e^{-\alpha}\alpha^{\alpha-1/2} \leq \Gamma(\alpha) \leq C_2\alpha^{\alpha-1/2}.$$

First,

$$\begin{aligned} \mathbb{P}(\tau > a) &= \mathbb{P}(X < 1/a) = c \int_0^{1/a} e^{-at}t^{a-1}dt \\ &\leq c \int_0^{1/a} t^{a-1}dt = \frac{1}{a} \frac{a^a}{\Gamma(a)} \frac{1}{a^a} = C \frac{1}{\sqrt{a}} \left(\frac{a}{e}\right)^a \\ &\leq e^{-Ca \log a}. \end{aligned}$$

The analysis of $\mathbb{P}(\tau \in [2a, 4a])$ follows similarly to the previous display, only the direction of the inequality needs reversal. To that end, lower-bound e^{-at} on $(0, 1/a)$ by e^{-1} and use the upper bound for $\Gamma(a)$.

Finally, consider $\mathbb{P}(\tau < 1/a)$. Note that for $t > a$, $t^{a-1} < e^{-at/2}$. Hence,

$$\begin{aligned} \mathbb{P}(\tau < 1/a) &= \mathbb{P}(X > a) = c \int_a^\infty e^{-at}t^{a-1}dt \\ &\leq c \int_a^\infty e^{-at/2}dt = (2/a)c e^{-a^2/2} \end{aligned}$$

The result follows since $c \leq C\sqrt{a}e^a$. \square

LEMMA A.2. *If $\|\Lambda_{0n}\|_2 \in [\kappa_{1n}, \kappa_{2n}]$, $\kappa_{2n} > 1$ and $\lambda_j \sim N(0, 1)$, $j = 1, \dots, p$ and $\sigma_n^2 \sim \text{IG}_{[0, M]}(a, b)$ and $\sigma_{0n}^2 \in [0, M]$, then for $0 < \epsilon < 1$,*

$$P(\|\Sigma_n - \Sigma_{0n}\|_F < \epsilon) \geq \exp\{-\kappa_{2n}^2 + p_n + p_n \log(\epsilon/2\kappa_{2n})\}.$$

PROOF. Since $\sigma_{0n}^2 \in [0, M]$ and $\sigma_n^2 \sim \text{IG}_{[0, M]}(a, b)$, it is enough to assume that $\sigma_n^2 = \sigma_{0n}^2$ to derive the prior concentration. Since $\Sigma_n = \Lambda_n \Lambda_n' + \sigma_{0n}^2 I_p$, we will first express the concentration of Σ_n around Σ_{0n} in terms of concentration of Λ_n around Λ_{0n} . Observe that

$$\begin{aligned} \|\Sigma_n - \Sigma_{0n}\|_F &\leq \|\Lambda_n \Lambda_n' - \Lambda_{0n} \Lambda_{0n}'\|_F \\ &\leq \|(\Lambda_n - \Lambda_{0n}) \Lambda_n'\|_F + \|\Lambda_{0n} (\Lambda_n - \Lambda_{0n})'\|_F \end{aligned}$$

Using the inequality $\|AB\|_F \leq \|A\|_2 \|B\|_F$, we have

$$\begin{aligned} \|(\Lambda_n - \Lambda_{0n})\Lambda_n'\|_F + \|\Lambda_{0n}(\Lambda_n - \Lambda_{0n})'\|_F &\leq \|\Lambda_n - \Lambda_{0n}\|_F (\|\Lambda_{0n}\|_2 + \|\Lambda_n - \Lambda_{0n}\|_F) \\ &\leq 2\kappa_{2n} \|\Lambda_n - \Lambda_{0n}\|_F. \end{aligned}$$

We estimate a lower bound for $\mathbb{P}\{\|\Lambda_n - \Lambda_{0n}\|_F < \epsilon/(2\kappa_{2n})\}$ below. Clearly $\Lambda_n \sim N_{p_n}(0, I_{p_n})$. Observe that the RKHS of p_n -dimensional Gaussian random vector Λ_n is the range $\{x : x \in \mathbb{R}^{p_n}\}$ equipped with the inner product $\langle x, y \rangle = x'y$. Since $\mathbb{P}\{\|\Lambda_n - \Lambda_{0n}\|_F < \epsilon/(2\kappa_{2n})\} = \mathbb{P}\{\|\Lambda_n - \Lambda_{0n}\|_2 < \epsilon/(2\kappa_{2n})\}$, we have by Borel's inequality for the p_n -dimensional Gaussian random vector λ_n

$$\mathbb{P}\{\|\Lambda_n - \Lambda_{0n}\|_F < \epsilon/(2\kappa_{2n})\} \geq \exp(-\|\Lambda_{0n}\|_2^2) \mathbb{P}\{\|\Lambda_n\|_2 < \epsilon/(2\kappa_{2n})\}.$$

Using the fact that $-\log(2\phi(x) - 1) < 1 + |\log x|$ for $0 < x < 1/2$, we have

$$\mathbb{P}\{\|\lambda_n\|_2 < \epsilon/(2\kappa_{2n})\} \geq \exp\{-p_n + p_n \log(\epsilon/(2\kappa_{2n}))\}.$$

The proof follows immediately. \square

Proof of Lemma 4.2. We begin with the observation

$$P(\|\eta - \eta_0\|_2 < \delta) \geq \prod_{j=1}^s P(|\eta_j - \eta_{0j}| < \delta/\sqrt{s}).$$

Now if $|\eta_{0j}| > \delta/\sqrt{s}$, then

$$\begin{aligned} P(|\eta_j - \eta_{0j}| < \delta/\sqrt{s}) &= \frac{1}{2} e^{-|\eta_{0j}|/\psi_j} \{e^{\delta/(\psi_j \sqrt{s})} - e^{-\delta/(\psi_j \sqrt{s})}\} \\ (A.1) \quad &\geq \frac{1}{2} e^{-|\eta_{0j}|/a} \{1 - e^{-\delta/(b\sqrt{s})}\}. \end{aligned}$$

On the other hand, if $|\eta_{0j}| \leq \delta/\sqrt{s}$, then observe that the interval $(\eta_{0j} - \delta/\sqrt{s}, \eta_{0j} + \delta/\sqrt{s})$ contains either $(0, \delta/\sqrt{s})$ or $(-\delta/\sqrt{s}, 0)$ depending on the sign (positive or negative respectively) of η_{0j} . Since a $DE(\psi_j)$ density is symmetric about the origin, each of the two intervals have the same probability implying

$$\begin{aligned} P(|\eta_j - \eta_{0j}| < \delta/\sqrt{s}) &\geq \frac{1}{2} \{1 - e^{-\delta/(\psi_j \sqrt{s})}\} \\ (A.2) \quad &\geq \frac{1}{2} \{1 - e^{-\delta/(b\sqrt{s})}\}. \end{aligned}$$

The desired inequality in Lemma 4.2 follows from combining (A.1) & (A.2).

Proof of Lemma 4.4. Using $\Gamma(x+1) = x\Gamma(x)$, note that $g(x) = -\log\{\Gamma(x+1)\}/x$. Hence, $\lim_{x \rightarrow 0} g(x) = \frac{d}{dx}[-\log\{\Gamma(x+1)\}]_{x=1} = \gamma_0$. We shall now show that $h(x) = -g(x)$ is increasing on $(0, 1/2)$. To that end, $h'(x) = \{x\Psi(x+1) - \log\Gamma(x+1)\}/x^2$, where $\Psi(x) = d/dx[\log\Gamma(x)]$ is the digamma function. Let $h_1(x) = x^2h'(x)$. Clearly, $h_1(0) = 0$. Further, $h_1'(x) = x\Gamma''(x+1) > 0$ since the gamma function is convex. Thus $h_1(x) > 0$ implying $h(x) > 0$ on $(0, 1/2)$, which concludes the proof.

Proof of Lemma 6.1. First observe that to prove Lemma 6.1, it is enough to show that $\mathcal{D}_n \geq e^{-n\delta_n^2}$ for a probability measure Π_n on $\{\Sigma_n : \|\Sigma_n - \Sigma_{0n}\|_F < \delta_n\}$.

By Jensen's inequality,

$$\log \mathcal{D}_n \geq \int \left[\frac{n}{2} \log |\Sigma_{0n} \Sigma_n^{-1}| - \frac{1}{2} \sum_{i=1}^n y_i^T (\Sigma_n^{-1} - \Sigma_{0n}^{-1}) y_i \right] \Pi_n(d\Sigma_n).$$

Letting $Q_i = y_i^T (\Sigma_n^{-1} - \Sigma_{0n}^{-1}) y_i$, one clearly has $\mathbb{E}_{\Sigma_{0n}} Q_i = \text{tr}(\Sigma_{0n} \Sigma_n^{-1} - I_p)$. Let $W_i = Q_i - \text{tr}(\Sigma_{0n} \Sigma_n^{-1} - I_p)$ and $S_n = \sum_{i=1}^n W_i$. We first show:

LEMMA A.3. *If $\|\Sigma_n - \Sigma_{0n}\|_F \leq \delta_n$ with $\delta_n/s_{\min}(\Sigma_{0n}) \rightarrow 0$ as $n \rightarrow \infty$, then for sufficiently large n ,*

$$\log |\Sigma_{0n} \Sigma_n^{-1}| - \text{tr}(\Sigma_{0n} \Sigma_n^{-1} - I_{p_n}) \geq -C \frac{\log(\varrho_n) \delta_n^2}{s_{\min}(\Sigma_{0n})^2},$$

for some absolute constant $C > 0$ and $\varrho_n = s_{\max}(\Sigma_{0n})/s_{\min}(\Sigma_{0n})$.

PROOF. Let H_n denote the symmetric positive definite matrix $\Sigma_n^{-1/2} \Sigma_{0n} \Sigma_n^{-1/2}$ with eigenvalues $\psi_j > 0, j = 1, \dots, p_n$. Clearly,

$$\begin{aligned} & \log |\Sigma_{0n} \Sigma_n^{-1}| - \text{tr}(\Sigma_{0n} \Sigma_n^{-1} - I_{p_n}) \\ (A.3) \quad &= \log |H_n| - \text{tr}(H_n - I_{p_n}) = \sum_{j=1}^{p_n} [\log \psi_j - (\psi_j - 1)]. \end{aligned}$$

Consider the function $h_\beta(x) = \log x - (x-1) + \beta(x-1)^2/2$ on $(0, \infty)$ for $\beta > 1$. Clearly, $h_\beta(1) = 0$, $h'_\beta(x) = (x-1)(\beta - 1/x)$ and $h''_\beta(x) = \beta - 1/x^2$. Thus h_β has a local minima at $x = 1$ and is monotonically increasing on $(1, \infty)$. Moreover, h_β has a local maximum at $1/\beta$ and the function is monotonically increasing on $(0, 1/\beta)$ and monotonically decreasing on $(1/\beta, 1)$. Since $h(1) = 0$, this implies $h_\beta(1/\beta) > 0$ and h_β has the property that if $h_\beta(x^*) > 0$, then $h_\beta(x) > 0$ for all $x > x^*$. Now suppose $\varepsilon \in (0, 1/2)$. We shall show that

$h_\beta(x) \geq 0$ for all $x \geq \varepsilon$ if $\beta = 8 \log(1/\varepsilon)$. Based on the above discussion, it suffices to show that $h_\beta(\varepsilon) > 0$, which follows since $h_\beta(\varepsilon) = \log(1/\varepsilon)[4(1 - \varepsilon)^2 - 1] + (1 - \varepsilon) > 0$.

Using (iii) in Lemma 5.1, $s_{\min}(H_n) \geq s_{\min}(\Sigma_{0n})/\|\Sigma_n\|_2$. Further, since $\|\Sigma_n - \Sigma_{0n}\|_2 \leq \delta_n$, $\|\Sigma_n\|_2 \leq \|\Sigma_{0n}\|_2 + \delta_n \leq 2\|\Sigma_{0n}\|_2$ since $\delta_n \in (0, 1)$. Thus, choosing $\varepsilon_n = s_{\min}(\Sigma_{0n})/\{2s_{\max}(\Sigma_{0n})\}$, $\psi_j \geq \varepsilon_n$ for all $j = 1, \dots, p_n$ and the analysis in the preceding paragraph shows $\log \psi_j - (\psi_j - 1) \geq -C \log(1/\varepsilon_n)(\psi_j - 1)^2$. Using (R1) in Lemma 5.2, we thus obtain that the quantity in (A.3) is bounded below by $-C \log(1/\varepsilon_n) \|H_n - I_{p_n}\|_F^2$.

Further, by (i) in Lemma 5.1, $\|H_n - I_{p_n}\|_F^2 \leq \|\Sigma_n - \Sigma_{0n}\|_F^2 / \{s_{\min}(\Sigma_n)^2\}$. We next proceed to lower-bound $s_{\min}(\Sigma_n)$. Using $\|AB\|_2 \geq s_{\min}(A)\|B\|_2$ from (ii) in Lemma 5.1, $\|\Sigma_n - \Sigma_{0n}\|_2 \geq s_{\min}(\Sigma_{0n})\|\Sigma_{0n}^{-1}\Sigma_n - I_{p_n}\|_2$, implying $\|\Sigma_{0n}^{-1}\Sigma_n - I_{p_n}\|_2 \leq \delta_n/s_{\min}(\Sigma_{0n})$. Since $\delta_n/s_{\min}(\Sigma_{0n}) < 1$ by assumption, and the singular values of $\Sigma_{0n}^{-1}\Sigma_n - I_{p_n}$ are $|1/\psi_j - 1|$ by similarity, it follows that $s_{\min}(\Sigma_{0n}^{-1}\Sigma_n) > 1 - \delta_n/s_{\min}(\Sigma_{0n})$. Invoking (iii) of Lemma 5.1, we finally get $s_{\min}(\Sigma_n) > s_{\min}(\Sigma_{0n})\{1 - \delta_n/s_{\min}(\Sigma_{0n})\} > s_{\min}(\Sigma_{0n})/2$ for n sufficiently large. \square

Using Lemma A.3, one has

$$\begin{aligned} \log \mathcal{D}_n &\geq \int \frac{n}{2} \left[\log |\Sigma_{0n} \Sigma_n^{-1}| - \text{tr}(\Sigma_{0n} \Sigma_n^{-1} - I_p) - \frac{1}{n} \sum_{i=1}^n W_i \right] \Pi_n(d\Sigma_n) \\ &\geq \frac{1}{2} \left[-Cn\delta_n^2 \frac{\log \varrho_n}{s_{\min}(\Sigma_{0n})^2} - \int S_n \Pi_n(d\Sigma_n) \right]. \end{aligned}$$

Set $A_n = \{\mathbf{y}^{(n)} : |S_n| \leq C\sqrt{n \log n} \delta_n/s_{\min}(\Sigma_{0n})\}$. Since $n\delta_n^2/s_{\min}(\Sigma_{0n})^2 \rightarrow \infty$, $n\delta_n^2/s_{\min}(\Sigma_{0n})^2 > \sqrt{n} \delta_n/s_{\min}(\Sigma_{0n})$ for n large. Hence, on A_n , $D_n \geq e^{-Cn\delta_n^2/s_{\min}(\Sigma_{0n})^2}$. It thus remains to show that $\mathbb{P}_{0n}(A_n^c) \rightarrow 0$. To that end, let $\zeta_{0n}^2 = \mathbb{E}_{\Sigma_{0n}}(W_i^2)$. By a standard result on quadratic forms, $\zeta_{0n}^2 = \|H_n - I_{p_n}\|_F^2 \leq C\delta_n^2/s_{\min}(\Sigma_{0n})^2$ by previous calculations. The proof is completed by an application of Markov's inequality:

$$\mathbb{P}_{\Sigma_{0n}}(A_n^c) = \mathbb{P}_{\Sigma_{0n}}\{S_n^2 > C \log(n) n \delta_n^2/s_{\min}(\Sigma_{0n})^2\} \leq \frac{1}{C \log n}.$$

Proof of Lemma 6.2. With $B_n = \{|\text{supp}_{\delta'_n}(\Lambda_n)| > H \log p_n\}$ and using Lemma 6.1,

$$\begin{aligned} \mathbb{E}_{\Sigma_{0n}} [\Pi_n(B_n \mid \mathbf{y}^{(n)}) 1_{A_n}] &= \mathbb{E}_{\Sigma_{0n}} \left\{ \frac{\int_{B_n} \prod_{i=1}^n \frac{f_{\Sigma_n}(y_i)}{f_{\Sigma_{0n}}(y_i)} d\Pi_n(\Sigma_n)}{\int \prod_{i=1}^n \frac{f_{\Sigma_n}(y_i)}{f_{\Sigma_{0n}}(y_i)} d\Pi_n(\Sigma_n)} 1_{A_n} \right\} \\ &\leq \mathbb{E}_{\Sigma_{0n}} \left\{ \frac{\int_{B_n} \prod_{i=1}^n \frac{f_{\Sigma_n}(y_i)}{f_{\Sigma_{0n}}(y_i)} d\Pi_n(\Sigma_n)}{e^{-n\delta_n^2/s_{\min}(\Sigma_{0n})^2} \mathbb{P}(\|\Sigma_n - \Sigma_{0n}\|_F \leq \delta_n)} \right\} \\ &= \frac{\Pi_n(B_n)}{e^{-n\delta_n^2/s_{\min}(\Sigma_{0n})^2} \mathbb{P}(\|\Sigma_n - \Sigma_{0n}\|_F \leq \delta_n)}. \end{aligned}$$

Using Lemma 4.3, $\Pi_n(B_n) \leq e^{-C \log p_n}$. By (A4), $s_{\min}(\Sigma_{0n})$ is bounded below by a constant and $\mathbb{P}(\|\Sigma_n - \Sigma_{0n}\|_F \leq \delta_n) \geq e^{-C \log p_n}$ by Lemma 6.5.

Proof of Lemma 6.3. The proof for the second part follows along the same lines as the in proof for Lemma 6.2, observing that for large t_n ,

$$\begin{aligned} P(\sigma^2 > t_n) &\leq \frac{b^a}{\Gamma(a)} \int_{t_n}^{\infty} e^{-bx} x^{a-1} dx \\ &\leq \frac{b^a}{\Gamma(a)} \int_{t_n}^{\infty} e^{-bx/2} dx \\ &\leq C e^{-C' t_n}. \end{aligned}$$

For the first part, note that $\|\Lambda_n\|_2 \leq C \|\text{vec}(\Lambda_n)\|_1$ and $\mathbb{P}(\|\text{vec}(\Lambda_n)\|_1 > t_n) \leq e^{-C \log p_n}$ by Lemma 4.5.

REFERENCES

- ARMAGAN, A., DUNSON, D. and LEE, J. (2011). Generalized double Pareto shrinkage. *Arxiv preprint arxiv:1104.0861*.
- ARMINGER, G. and MUTHÉN, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika* **63** 271–300.
- BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.
- BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.
- BARTHOLOMEW, D. J. (1987). *Latent Variable Models and Factor Analysis*. Oxford University Press, New York.
- BELITSER, E. and GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics* **31** 536–559.
- BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* **36** 2577–2604.

- BICKEL, P. J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36** 199–227.
- BONTEMPS, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics* **39** 2557–2584.
- CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.
- CAI, T. T., ZHANG, C. H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38** 2118–2144.
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480.
- CARVALHO, C., LUCAS, J., WANG, Q., NEVINS, J. and WEST, M. (2008). High-dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association* **103** 1438–1456.
- CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straws in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics*, to appear.
- EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* **36** 2717–2756.
- FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147** 186–197.
- FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics* **39** 3320–3356.
- FURRER, R. and BENGTTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis* **98** 227–255.
- GHOSAL, S. (1997). Normal approximation to the posterior distribution for generalized linear models with many covariates. *Mathematical Methods of Statistics* **6** 332–348.
- GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–531.
- HANS, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association* **106** 1383–1393.
- HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98.
- JIANG, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *The Annals of Statistics* **35** 1487–1511.
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29** 295–327.
- JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.
- KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics* **4** 1225–1257.
- LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37** 4254–4278.
- LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* **40** 694–726.
- LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *The Annals of Applied Statistics* **2** 245–263.
- LUCAS, J. E., CARVALHO, C., WANG, Q., BILD, A., NEVINS, J. R. and WEST, M. (2006). Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for*

- Gene Expression and Proteomics* (K. A. Do, P. Müller and M. Vannucci, eds.) 155–176. Cambridge University Press.
- MUIRHEAD, R. J. *Developments in eigenvalue estimation*. Advances in Multivariate Statistical Analysis (A.K. Gupta, ed.), Reidel, Dordrecht, 277–288.
- PARK, T. and CASELLA, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103** 681–686.
- POLSON, N. G. and SCOTT, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9* (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds.) 501–538. Oxford University Press, New York.
- ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.
- SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38** 2587–2619.
- SONG, X. Y. and LEE, S. Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical and Statistical Psychology* **54** 237–263.
- VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *Arxiv preprint arxiv:1011.3027*.
- WEST, M. (2003). Bayesian factor regression models in the “large p , small n ” paradigm. In *Bayesian Statistics 7* (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds.) 733–742. Oxford University Press, New York.
- WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844.
- WU, W. B. and POURAHMADI, M. (2010). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica* **19** 1755.
- ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15** 265–286.

BOX 90251, OLD CHEMISTRY BUILDING
DURHAM, NC 27708

E-MAIL: dp55@stat.duke.edu
ab179@stat.duke.edu; pillai@fas.harvard.edu; dunson@stat.duke.edu